Scientific Research Publishing

Journal of Transportation Technologies, 2019, 9, 109-119
http://www.scirp.org/journal/jtts
ISSN Online: 2160-0481
ISSN Print: 2160-0473

Case Study of Four Vehicle Reliability Comparison Based on Survival Analysis

Pengzhou Xu[1], Junhui Gao[2*]

[1]Jiangsu Tianyi High School, Wuxi, China
[2]American and European International Study Center, Wuxi, China
Email: *jhgao68@163.com

How to cite this paper: Xu, P.Z. and Gao, J.H. (2019) Case Study of Four Vehicle Reliability Comparison Based on Survival Analysis. *Journal of Transportation Technologies*, **9**, 109-119.
https://doi.org/10.4236/jtts.2019.91007

Received: December 19, 2018
Accepted: January 13, 2019
Published: January 16, 2019

Abstract

The paper is written to analyze the behavior of a selected set of vehicles with different makes, on how they survive after each owner change. The data come from Github. The four cars are Honda Accord, Mini Cooper, Chevy Cavalier, and Toyota Avalon. The two faults are Engine System and Transmission System. The data are from 1996 to 2012. The paper used the Kaplan-Meier curve to survival analysis; the paper also calculates and discusses the self-comparison of each car's four time periods, the four-stage failure rate through median comparison, and the median comparison of fault conditions in all years. We find that all the vehicle types have gotten better with the years and Toyota vehicles are more reliable than Honda.

Keywords

Survival Analysis, Automobile, Reliability, Case Study

1. Introduction

Vehicle survival is a concept concerning total time a vehicle works after it is sold to a customer and the malfunctions of the vehicle. Vehicle survival analysis is utilized in numerous areas like vehicle quality assessment. For example, people are able to anticipate latent problems that might occur on vehicles to ensure the driver and passengers' safety; survival analysis is also employed in large-scale vehicle scrappage programs to maximize the usage vehicles' abilities and maintain the price of vehicles.

Furthermore, the vehicle survival analysis can also be used to estimate the car stability even before customers purchase it. In this way, money is used more efficiently. This analysis related to specific vehicles provides car manufacturers with a great opportunity to make an improvement in their products, attracting new

consumers and ensure old consumers' loyalty.

Former researchers have done a similar analysis to estimate vehicle performances. Data mining and neural network methods are utilized to estimate the reliability of a vehicle [1], using MATLAB2007 to evaluate the engine performance. People also do such reliability assessment on race cars. Fault tree, finite element analysis are used to estimate the full car reliability of FSAE race cars [2]. Diesel engines are also researched by experts to find out its reliability. The methods integrate Weibull Model and Dempster-Shafer evidence theory to describe the regularities of distribution of its failure rate of each part of the engine when it still functions [3]. In this paper, we are going to analyze four types of vehicles to compare the overall performance of each vehicle and determine the car with the greatest development on reliability over years.

2. Data Sources

In this paper, we adopt data from Github (https://github.com/tcrug/car-reliability). Data is in the form of charts. In the data set, there are totally six columns in the data chart. The first column represents the date the vehicle was bought; the second column shows the vehicle manufacturer; the third column is the specific type of vehicle produced by the manufacturer; the fourth column lists out the total distance traveled by the vehicle before it was examined at it first malfunction; the fifth and sixth columns represents the state of engine and transmission system respectively. Four different types of vehicles from different car brands are analyzed, like Honda, Toyota, MINI, and Chevrolet. To make the data represent the complicated vehicle market more generally, we deliberately chose car manufacturers from three different countries which stand for different manufacture criteria and styles. The malfunctions are separated into two major categories: engine problem and transmission problem. The vehicle is examined at the time the state of these two parts are represented by 0 and 1. 0 means no problem found after the vehicle examination, whereas 1 represents that problem exists in corresponding part.

3. Related Technology

In this paper, we employ mainly three methods: non-parameter method, semi-parametric method and parametric method.

3.1. Non-Parametric Analysis

Kaplan-Meier estimation graph and log-rank test are utilized to analyze the state of each type of vehicles.

3.1.1. Kaplan-Meier Estimator

The graph of Kaplan-Meier estimator declines like stairs. It is composed of multiple horizontal lines and vertical lines to reveal the chance of individual to survive within a given time. It is described as survival function. It is mainly used in medical treatment to estimate the probability for patients' to survive under certain circumstances, but in this paper it serves as the main method to evaluate ve-

hicle survival possibilities. The utility of this method on vehicles is essential for the promotion of vehicle production with higher qualities [4].

The estimator has the basic function of

$$\hat{s}(t)=\prod_{i:t_i\le t}\left(1-\frac{d_i}{n_i}\right)$$

t_i is a time when at least one event happened, d_i is the number of events that happened at t_i. n_i is the individuals known to survive (have not yet had an event or been censored) at time t_i. There is no unknown parameter, so Kaplan-Meier can be include in non-parametric methods [5]. However, the d_i/n_i can be regarded as a parameter. We can use the method of maximum likelihood to estimate its value.

We hypothesize the new function to be

$$\hat{s}(t)=\prod_{i:t_i\le t}\left(1-h_i\right)$$

The likelihood function is

$$L\left(h_{j:j\le i}\right)=\prod_{j=1}^{i}h_j^{d_j}\left(1-h_j\right)^{n_j-d_j}$$

To maximize the likelihood function, just simplify the function using natural logarithm.

$$\ln(L)=\sum_{j=1}^{i}d_j\ln\left(h_j\right)+\left(n_j-d_j\right)\ln\left(1-h_j\right)$$

$$\frac{\partial\ln(L)}{\partial h_i}=\frac{d_i}{\hat{h}_i}-\frac{n_i-d_i}{1-\hat{h}_i}=0\Rightarrow\hat{h}_j$$

The Kaplan-Meier estimator is one of the most frequently used method for survival analysis. It has a comparably advantage in estimating the death rate which is the rate of malfunction in vehicles in each part. Also, the result is clearer since it is visualized.

3.1.2. Log-Rank Test

The logrank test statistic compares estimates of the hazard functions of the two groups at each observed event time. It is constructed by calculating the observed and expected number of events in one of the groups at each observation time and then add these estimates to obtain an overall summary throughout the focused period where there is an event [6].

Let $j = 1, ..., J$ be the distinct times of observed events in either group. For each time j, let N_{1j} and N_{2j} be the number of subjects "at risk" (have not yet had an event or been censored) at the start of period j. Let $N_j = N_{1j} + N_{2j}$. Let O is the observed number of events. The expectation value of the log-rank test is E_{ij}, the variance difference is V_j. [7]

$$Z=\frac{\sum_{j=1}^{J}\left(O_{1j}-E_{1j}\right)}{\sqrt{\sum_{j=1}^{J}V_j}}\sim N(0,1)$$

Calculated outcome should be tested using Z test above and determined whether it is in the acceptable range.

Log-rank test can estimate the difference between two groups with significantly different risks, but it is only a test for significance, so it will not be the primary resolution in this paper.

3.2. Semiparametric Analysis

COX Regression Model

COX regression model uses $h(t,X)=h_0(t)\exp(\beta_1 X_1+\cdots+\beta_m X_m)$ as a variable in the middle instead of directly determine the relationship between the causing factor X and the survival function $S(t,x)$ [8]. The main idea of COX regression model as a semiparametric method is that the parameter β in the model is able to be determined without knowing $h(t,X)$. There is a prerequisite for using the COX model: the effect of each factor X do not vary as time passes on. We decide the function of each factor by determine the relative risk between the exposed group and non-exposed group [9].

$$RR=\frac{h(t,X_i)}{h(t,X_j)}$$

Cox regression model takes multiple factors which will affect the studied subject's survival time.

3.3. Parametric Analysis

3.3.1. Weibull Distribution and Exponential Distribution

Exponential distribution and Weibull distribution measure the status of the occurrence of a specific event in a time interval. Exponential distribution and Weibull distribution has a probability density function respectively:

$$f(x)=\begin{cases}\lambda e^{-\lambda x}(x>0)\\ 0(x\le 0)\end{cases},\quad f(x;\lambda,k)=\begin{cases}\frac{k}{\lambda}\left(\frac{k}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}(x\ge 0)\\ 0(x<0)\end{cases}$$

Weibull distribution and exponential distribution are very alike [10]. When k = 1 in Weibull distribution, the density function becomes the same as that of exponential distribution.

3.3.2. Parameter Estimation

The parameters have to be determined clearly to draw the precise probability density function. We mainly use two methods for parameter estimation-point estimation and the maximum likelihood estimation [11]. Since the maximum likelihood method has more accuracy, we mainly focus on this method to determine the parameters.

Parametric estimation can be combined with predefined equations and functions to estimate duration of a project.

3.3.3. Exponential Regression and Weibull Regression

To determine whether there is significant cause and effect relationship, we have

to do regression test to the outcome from Weibull and Exponential distribution. First, we hypothesize that there is no relationship between the factor and the result, then we propose the formula and plug in the required data presented in the data set. Next, we determine the rejection region and see if the value falls within this range. Finally, we give the result whether accept the hypothesis or not.

4. Methods

4.1. Overall Description of Data

Use SQL to sort out the data distribution of the four cars, listed in Table 1. Numbers in the table indicate how many pieces of failure data are available for a particular type of vehicle in the corresponding year. For example, Honda had 393 samples in 1996 and MINI had 129 samples in 2002.

As we can see from Table 1, the data is from 1996 to 2012. Among them, Honda and Toyota have data in all years, while Chevrolet lacks data from 2006 to 2012. MINI lacks data from 1996 to 2001.

4.2. Analysis Strategy

We adopt non-parametric method to analyze the faults of automobiles. Based on the data distribution, considering the lack of data, we are ready to analyze and model the data from two angles. The first angle is to compare the survival curves of the four cars, using the Kaplan-Meier estimator. The second angle is to compare the survival curves of different time periods. We divide the time into four sections, 1996-1999, 2000-2003, 2004-2007, 2008-2012. The last stage is one year longer than the first three stages, considering that the data for 2012 is relatively small.

Table 1. Time distribution of available samples of four cars.

	Chevrolet	Honda	MINI	Toyota
1996	48	393		109
1997	98	509		139
1998	116	977		171
1999	146	1171		152
2000	209	1382		354
2001	222	1271		229
2002	335	1419	129	175
2003	348	1650	254	135
2004	354	1152	239	96
2005	195	984	337	123
2006		574	299	203
2007		570	245	105
2008		481	194	49
2009		231	158	9
2010		206	61	11
2011		101	25	11
2012		47	12	1

4.3. Programming Tools

The programming language uses python 3.6 [12] and pandas [13], and the survival analysis uses KM related functions in third-party package lifetime [14]: Kaplan Meier Fitter, multivariate logrank test.

5 Results

5.1. Compare the Survival Curves of Four Cars

Comparing four types of vehicles in all years by K-M method, the results of calculating the two kinds of faults are shown in **Figure 1** and **Figure 2**, respectively.

From **Figure 1**, we find that almost all cars in Honda had a fault before the mileage of 400,000 miles. The MINI car, before 200,000 miles, had a general car failure. The failure of the other two cars is significantly better than Honda and MINI.

From **Figure 2** we find that Toyota's fault condition is significantly better than the other three cars.

5.2. Compare Survival Curves over Different Time Periods

Different time periods, K-M overall comparison, two faults, the results are shown in **Figure 3**, **Figure 4**.

From **Figure 3**, we find that for the first type of failure, 2008-2012 is the best period of reliability (low failure rate), followed by 2004-2007, then 1996-1999 and finally 2000-2003. This shows that the fault changes in a good direction over time.

From **Figure 4**, we find that for the second type of failure, the failure situation in 2004-2007 is the best, followed by 2008-2012, followed by 1996-1999, and finally 2000-2003. This shows that the fault changes in a good direction over time, but the fluctuation is larger than the fault.

6. Discussion

In the fifth part of the article, we give two results, which are the survival curves of the four cars, and the survival curves of different time periods. In order to

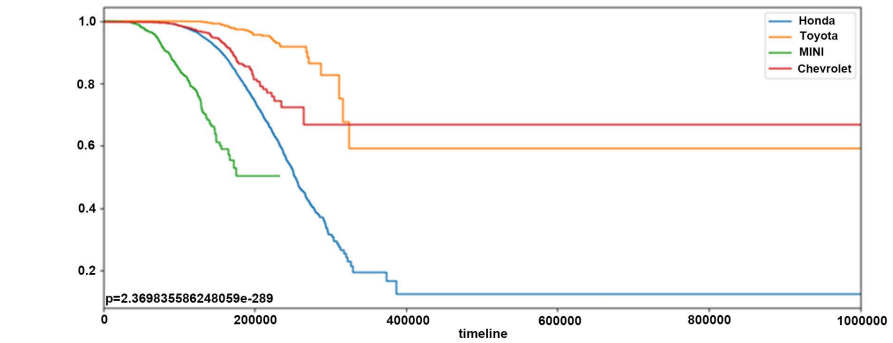

Figure 1. Comparison of survival curves of the first type of fault.

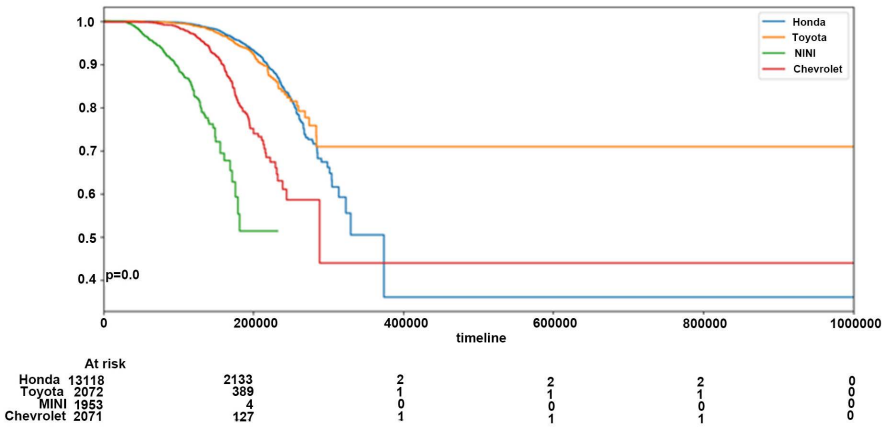

Figure 2. Comparison of survival curves of the second type of fault.

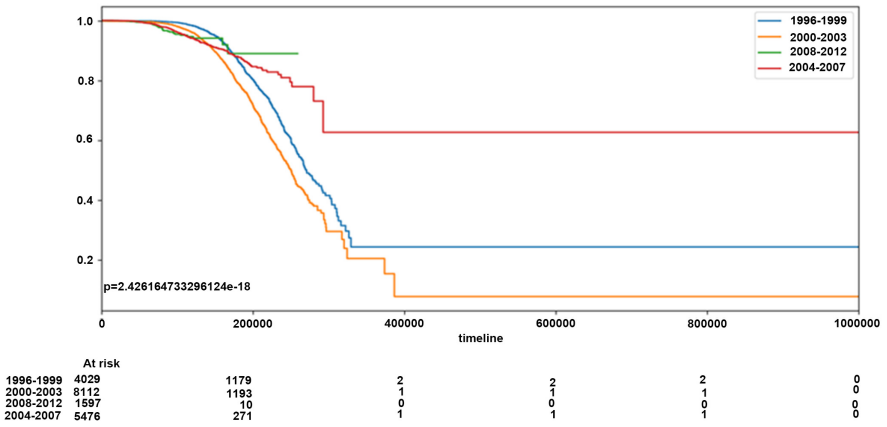

Figure 3. Comparison of survival curves of the first fault in four time periods.

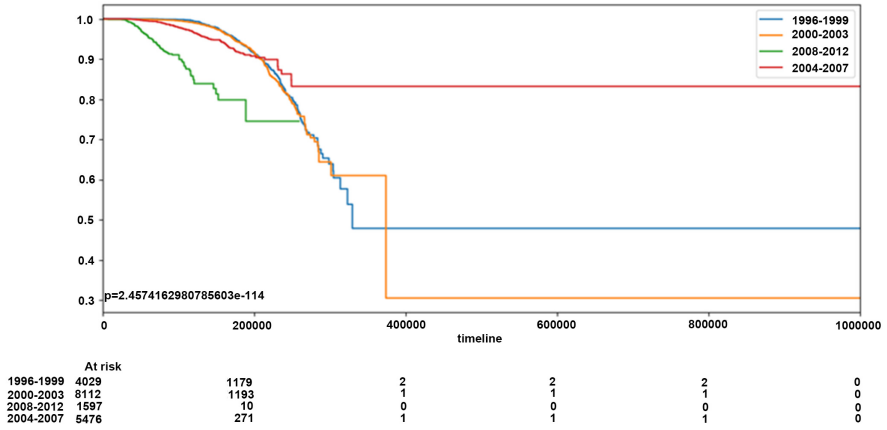

Figure 4. Comparison of survival curves of the second fault in four time periods.

compare the reliability of the four cars more deeply, we will continue to carry out some analysis and calculation here.

6.1. Self-Comparison of Each Car in Four Time Periods

Here we select Honda and Toyota, which compare the faults of each of the four

time periods.

From Figure 5, we find that for the second fault of Honda, the two phases of 2004-2007 and 2008-2012 are relatively close, which are better than the other two time periods.

From Figure 6, we find that for the second failure of Toyota, the two phases of 2004-2007 and 2008-2012 are relatively close, and the two phases of 2000-2003 and 1996-1999 are similar Closer.

6.2. Four-Stage Failure Rate by Median Comparison

The median here is the average mileage of all cars when 50% of cars fail. We compare the median of the four stages here, and the results are shown in Table 2.

From Table 2, we can find that in the four time periods, fault 1 has a median in the two stages of 1996-1999 and 2000-2003, and the other two stages do not. This means that in the two periods of 2004-2007, 2008-2012, the faulty car has never reached 50%, so overall, from 1996 to 2012, the fault 1 situation is changing in a good direction. Although the data of 2000-2003 is poor with the data of 1996-1999. The situation of fault 2 is basically similar to that of fault 1, except that in the first two phases, the median mileage of fault 2 is significantly more than that of fault 1.

6.3. Compare the Failures of All Years by Median

Similar to the method of 6.2, we compare the median of all years here, and the calculation results are listed in Table 3. Considering that only Honda and Toyota have data for all years, the other two models are not counted.

In Table 3, after 2002, the median of the two faults of Honda's car could not be calculated, indicating that the reliability of the car has improved since 2002. In the whole 16 years, Toyota has only the faults of 1998 and 2000, and the median of fault 2 in 2009 can be calculated. Overall, Toyota's reliability is better than Honda.

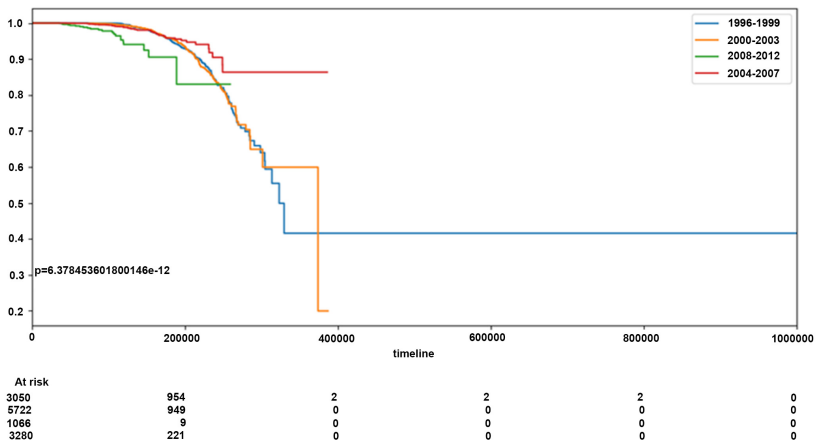

Figure 5. Comparison of survival curves of the second fault in four time periods.

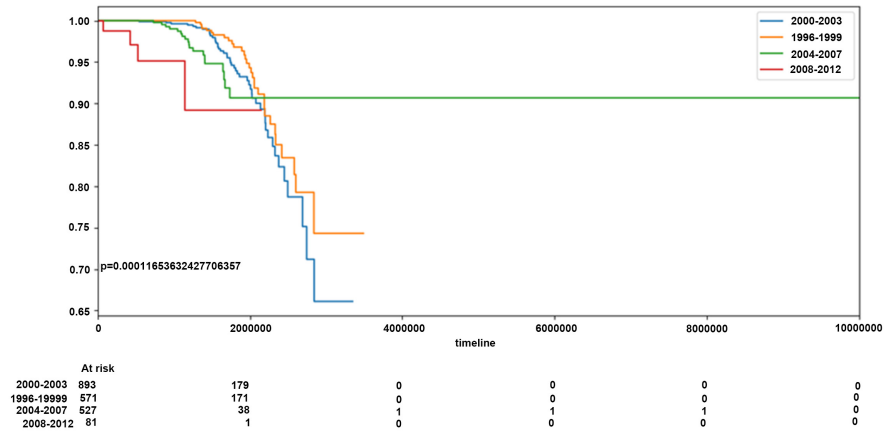

Figure 6. Comparison of survival curves of the second failure of toyota in four time periods.

Table 2. Median comparison of the four stages.

	the first fault	the second fault
1996-1999	270,919	329,885
2000-2003	250,720	374,217
2004-2007	inf	Inf
2008-2012	inf	inf

Table 3. Median comparison of four stages.

	Honda		Toyota	
	the first fault	the second fault	the first fault	the second fault
1996	296188	inf	inf	inf
1997	255692	inf	inf	inf
1998	260828	329,885	287496	inf
1999	257376	323,454	inf	inf
2000	237485	301,751	324,786	inf
2001	225687	inf	inf	inf
2002	212010	inf	inf	inf
2003	inf	inf	inf	inf
2004	inf	inf	inf	inf
2005	inf	inf	inf	inf
2006	inf	inf	inf	inf
2007	inf	inf	inf	inf
2008	inf	inf	inf	inf
2009	inf	inf	inf	114,932
2010	inf	inf	inf	inf
2011	inf	inf	inf	inf
2012	inf	inf	inf	inf

6.4. Inadequacies of Research

The most obvious shortcoming of this study is that the data source is single, and the data of other two cars is incomplete. In 17 years, there are 6 years of missing data.

7. Conclusion

This paper analyzes the survival of engine and transmission faults and compares the reliability of four vehicles from different manufacturers. The research in this paper shows that by applying the Kaplan-Meier fitter method and the log-rank test, we can not only get the most insight into improving the car brand, but also get the best performance. A comparative analysis of the four time periods suggests that the entire industry may be getting better. Data analysis can provide customers with very useful vehicle reliability information for their reference at the time of purchase. Survival analysis methods can also be applied to specific parts of a vehicle, such as the most common damaged parts on a vehicle—a tire or suspension. This aspect is also one of our follow-up studies.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Meng, S.P., Luo, H.Y. and Li, S. (2007) "Metal Heat Treatment" Automobile Reliability Analysis Method Based on Data Mining.

[2] Wang, J. (2015) The Reliability Analysis and Application on the FSAE Racing Vehicle. Doctoral Thesis, Hefei University of Technology, Hefei.

[3] Yang, Z.Q. (2015) Research on Reliability Analysis Data on Diesel Engines. Doctoral Thesis, Jiangxi University of Science and Technology, Ganzhou.

[4] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 457-481. https://doi.org/10.1080/01621459.1958.10501452

[5] Kaplan, E.L. (1983) This Week's Citation Classic. *Current Contents*, **24**, 14.

[6] Mantel, N. (1966) Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. *Cancer Chemotherapy Reports*.

[7] Peto, R. and Peto, J. (1972) Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society*, *Series A*, Blackwell Publishing, **135**, 185-207. https://doi.org/10.2307/2344317

[8] Cox, D.R. (1972) Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society Series B*. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[9] Cox, D.R. and Oakes, D. (1984) Analysis of Survival Data. Chapman & Hall, New York.

[10] Fréchet, M. (1927) Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathematique*, Cracovie.

[11] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994) Continuous Univariate Distributions. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*, 2nd Edition, John Wiley & Sons, New York.

[12] Python 3.6. https://www.python.org/downloads/release/python-360/

[13] Pandas. https://pandas.pydata.org/

[14] Lifelines. https://lifelines.readthedocs.io/en/latest/