Case Study 2, Pt. 1

Nathaniel Brown, In Hee Ho, Sarah Zimmermann

September 28, 2017

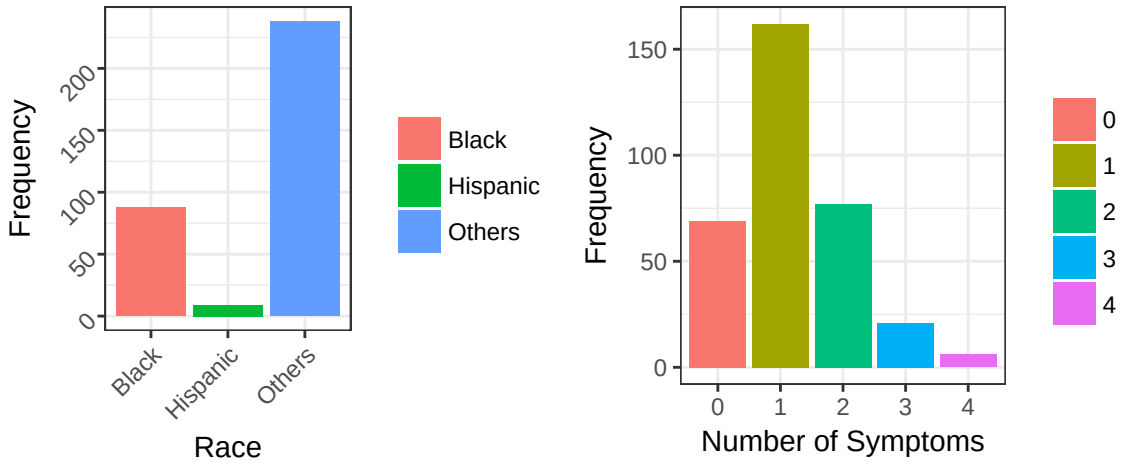
Introduction

The data are from a study of time to critical neurological assessment for patients with stroke-like symptoms who are admitted to the emergency room. We are interested in the factors predictive of the time to assessment following admission to the ED for n=335 patients with mild to moderate motor impairment. The goal of the analysis is to perform inferences on the impact of clinical presentation, gender, and race on time to neurological assessment, where clinical presentation is measured as the number of the four major stroke symptoms: headache, loss of motor skills or weakness, trouble talking or understanding, and vision problems.

Variable Exploration

The original data set contains 335 observations across 9 variables. They are defined as:

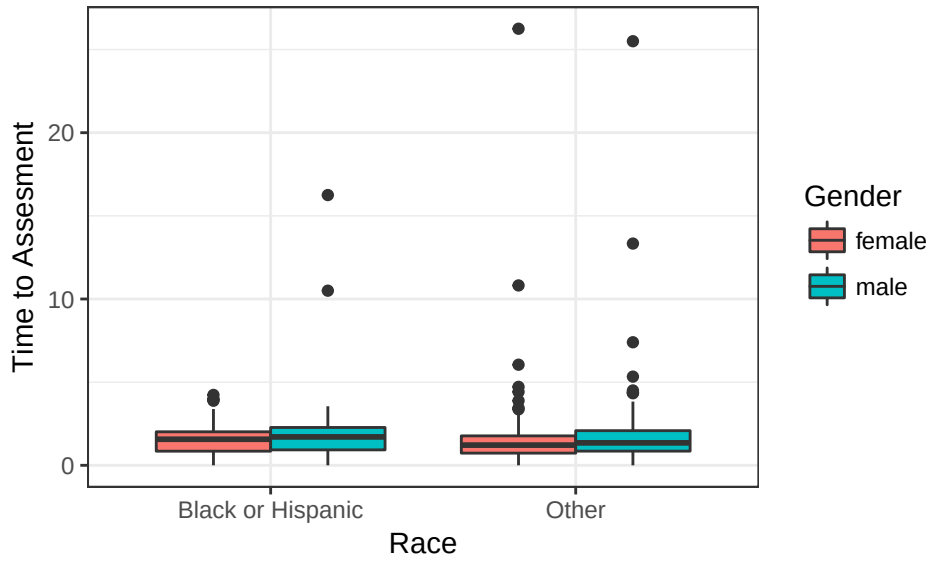
Variable Name	Short Description	Type
nctdel	min of neurologist time to assessment & CT scan from arrival at ER	continous
fail	1 if got neurologist/CT scan & 0 otherwise	categorical
male	1 if male, 0 if female	categorical
black	1 if black, 0 if not black	categorical
hisp	1 if hispanic, 0 if not hispanic	categorical
sn1	0/1 indicator 1 main symptom	categorical
sn2	0/1 indicator 2 main symptoms	categorical
sn3	0/1 indicator 3 main symptoms	categorical
all4	0/1 indicator all main sumptoms	categorical



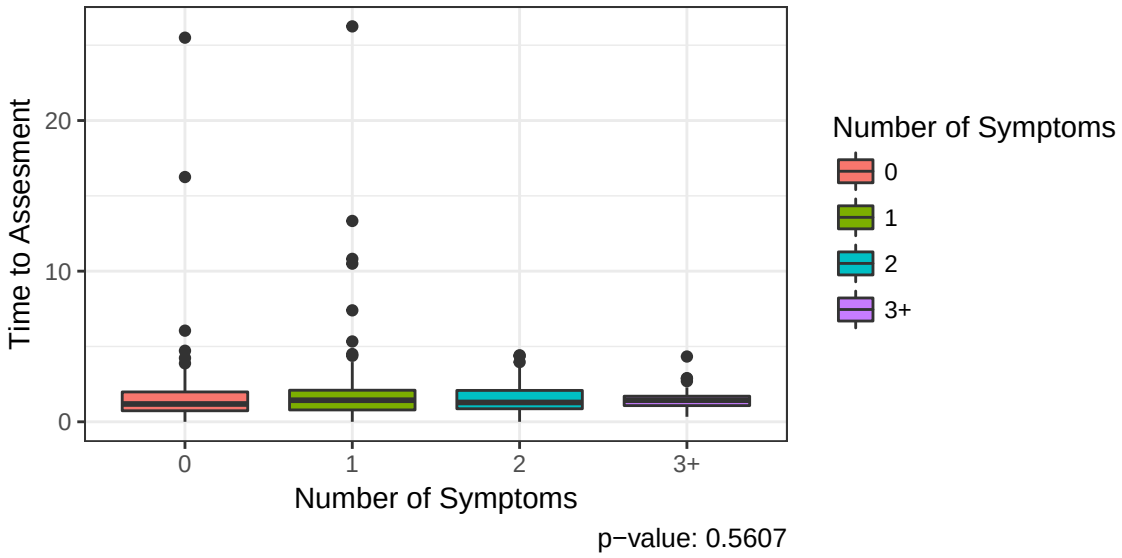
There are no missing values or apparently out of range values in our data, and therefore we had no need to clean the data. However, we created variables to better visualize the data. Race is a 2-level variable (Black/Hispanic and Other) which represents all those in the dataset who are Black or Hispanic in one category and non-Black and non-Hispanic in the other. We decided to group Black and Hispanic together because Blacks and Hispanics had relatively small sample sizes, constituting 97 out of 335 patients observed. Particularly, there were only 9 Hispanics, which is a population too small to represent the group and therefore difficult to draw conclusions upon. Next, symptom is a 4 level variable: "0" for those who had no symptoms, "1" for those who had 1 symptom, "2" for those who had 2 symptoms, and "3+" for those who had 3 or more symptoms. We grouped those patients which 3 or 4 symptoms together because there were only 6 individuals out of 335 observations in the dataset who had 4 symptoms, which is very small a population size to draw any conclusions on.

Exploratory Data Analysis

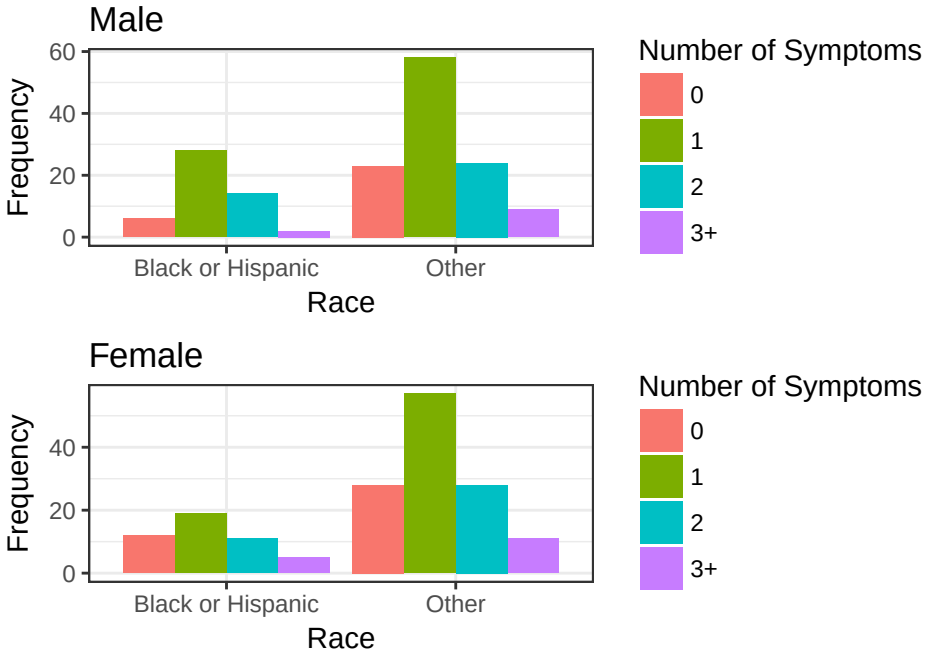
Data Visualizations



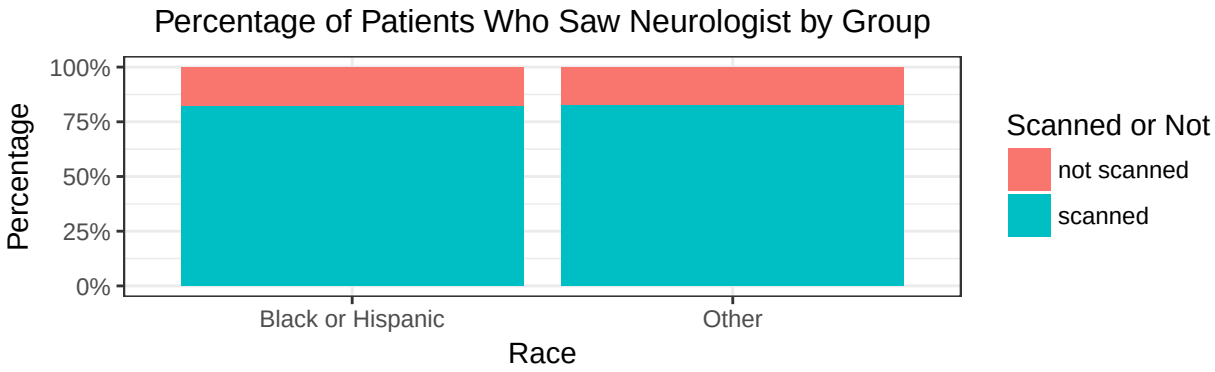
There is no visible difference in the median times to treatment between ethnic and gender groups. However, we see that non-Black and non-Hispanic group have higher variance compared to the Black or Hispanic group.

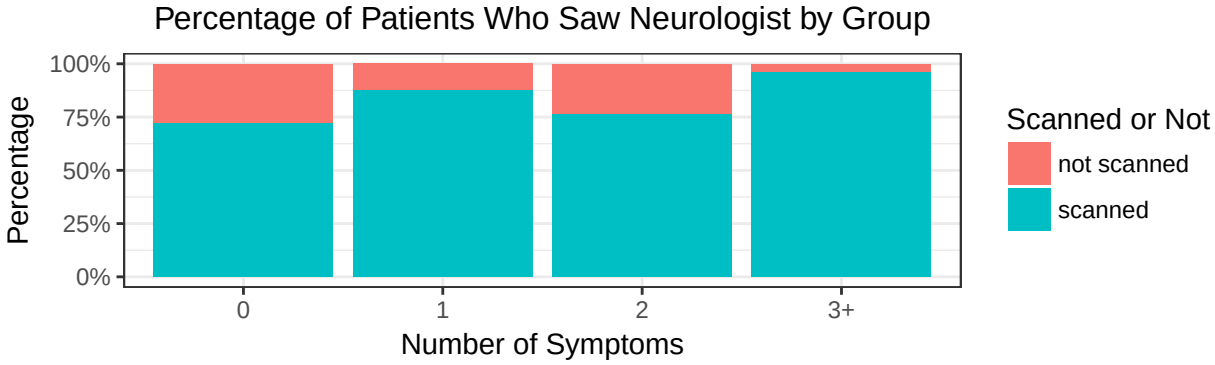


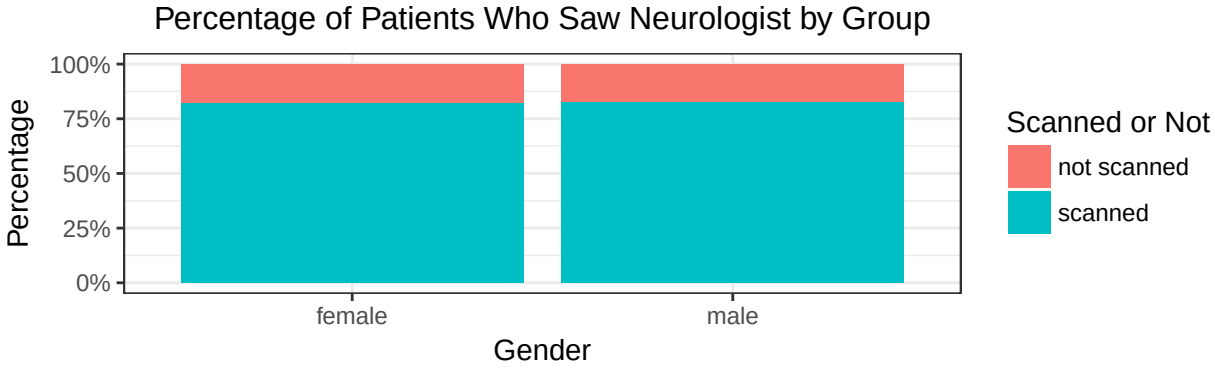
These boxplots illustrate the different distributions of time to treatment by the number of symptoms one shows. The median wait times are similar, as well as the means, according to an ANOVA test for equal means between the four groups (p=0.5607). However, the quantiles show that the distribution tends to be concentrated for the group with more symptoms.



We now examine the distribution of symptoms between genders and ethnicities. People, regardless of ethnicity, are most likely to arrive with one symptom than any other number, and the number of symptoms is right skewed. There is no noticeable difference in the distributions between the males and females.







Lastly, we observe the proportion of people from each group who got to see a neurologist or got a CT scan and who did not. We see that the proportion of the patients who got CT scan is similar across race and gender, while the patients who showed 3 or more major stroke symptoms are more likely to get a CT scan than patients with fewer stroke symptoms. It is also interesting to note that with 1 symptom are more likely to receive a scan than patients with 2.

Kaplan-Meier Analysis for Race, Gender, and Clinical Presentation

Background:

The Kaplan-Meier estimate is one of the best options to use to measure the fraction of subjects surviving for a certain amount of time. The curves are traditionally used to measure time to death, but the interpretation in this case is time to evaluation. The Kaplan-Meier estimate is also called as "product limit estimate". It involves computing probabilities of occurrence of an event at a certain point of time. Specifically, for each time interval, survival probability is calculated as the number of subjects surviving divided by the number of patients at risk. Subjects who have died, dropped out, or move out are not counted as "at risk" i.e., subjects who are lost are considered "censored" and are not counted in the denominator. Total probability of survival

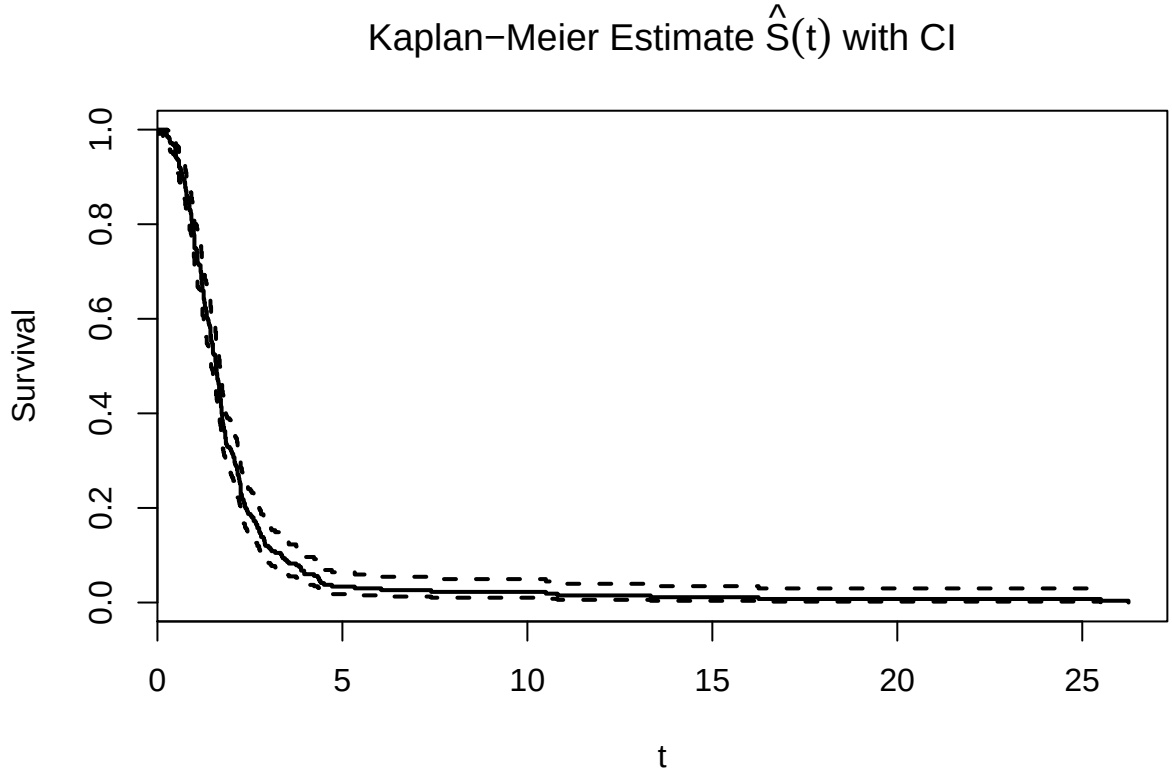
until that time interval is calculated by multiplying all the probabilities of survival at all time intervals preceding that time (by applying law of multiplication of probability to calculate cumulative probability). There are three assumptions used in this analysis. Firstly, we assume that at any time patients who are censored have the same survival prospects as those who continue to be followed. Secondly, we assume that the survival probabilities are the same for subjects recruited early and late in the study. Thirdly, we assume that the event happens at the time specified.

Using our graphs:

We can compare curves for different groups of subjects, such as the survival pattern for subjects on a standard therapy with those on a newer therapy. We can look for gaps in these curves in a horizontal or vertical direction. A vertical gap means that at a specific time point, one group had a greater fraction of subjects surviving. A horizontal gap means that it took longer for one group to experience a certain fraction of deaths. The two survival curves can be compared statistically by testing the null hypothesis i.e. there is no difference regarding survival among two classifications of population.

Analysis:

The first graph we created below is the overall survival curve. This curve shows how long the entire population in our dataset no matter the race, gender, or clinical condition will wait for evaluation. The graph also includes an upper and lower 95% confidence interval. Using this graph we know how the population is expected to wait. We can compare this estimated survival curve to the separate estimated survival curve for the different gender, race, and clincial condition groupings.

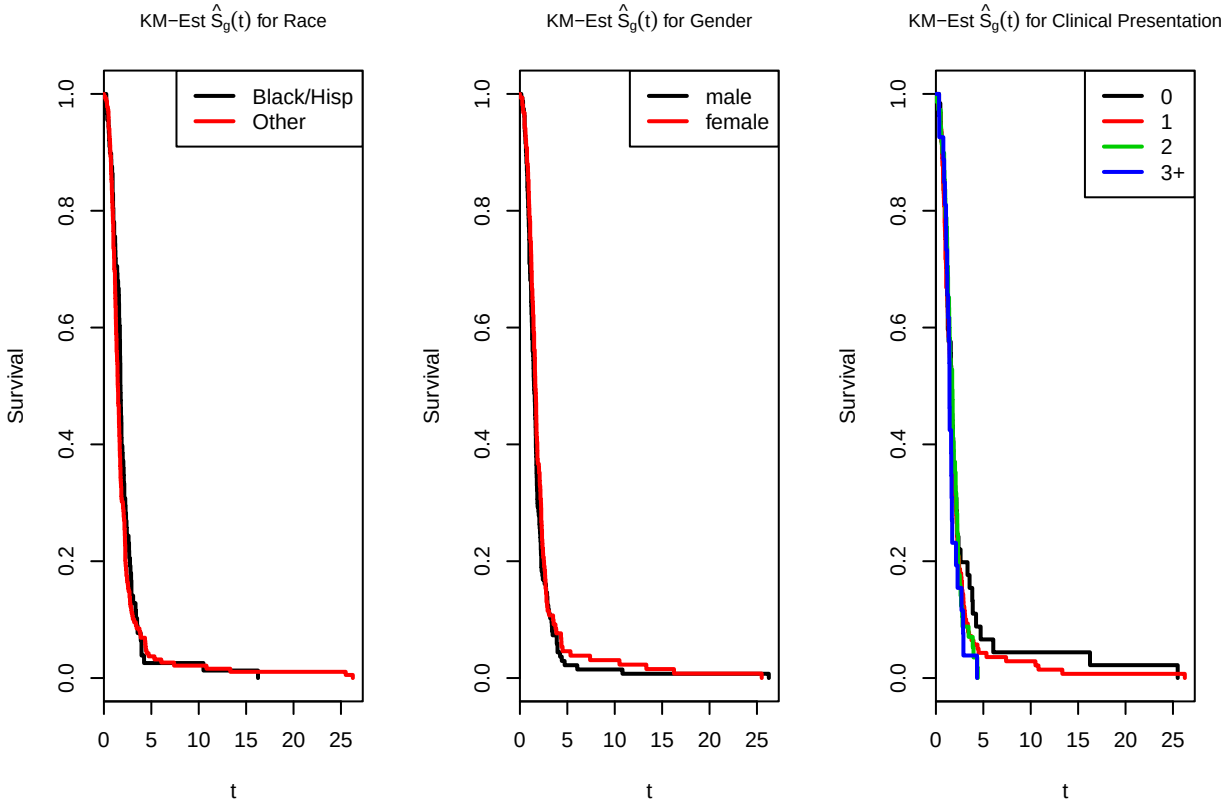


Below we see 3 estimate curves for gender (males vs. female), race (Black and Hispanic vs. Other), and clinical presentation (0 symptoms, 1 symptom, 2 symptoms, 3 or more symptoms).

Looking at our Kaplan-Meier estimate for Race, we see the survival curves are generally similar in the 0-5 time range. There is a steep decline in percent of the population waiting for an evaluation. After time 5, the survival curves diverge slightly, and sometimes there is a higher percentage of Black/Hispanic people waiting compared to Non-Black/Non-Hispanics and vice versa. The most interesting part of the comparison of curves is when the populations hit 0%. The Black/Hispanic group hits 0% between time 15 and 20, while the Non-Black/Non-Hispanic group hits 0% after time 25. This is something to look into in future analysis.

Next, looking at our Kaplan-Meier estimate for Gender, we see a very similar estimation of the survival curve between time 0 and 5. Also, the populations both hit 0% waiting at around the same time after time 25. It is interesting to note though after time 5, a higher percentage of females are waiting for an assessment than males; this is something to investigate in more detail in the future.

Our final graph is the Kaplan-Meier estimate for Clinical Presentation which is the number of symptoms a patient presents. There is difference between the estimated survival curves. We observe that people showing more symptoms have to wait a shorter amount of time for treatment to occur. It is interesting to see the 1 symptom group has all of their observations receive treatment slightly after the 0 symptom group.



In summary, in our EDA about Kaplan Meier estimates, we stated our null hypothesis is the survival curves for different populations will be the same. From our intial explorations we think we will reject the null hypothesis in favor of the alternative that the survival curves will be different for different classifications of a population i.e. Race, Gender, and Clinical Presentation have an effecct on getting a CT scan.

Approach for analyzing

For further analysis next week, our null hypotheses can be statistically tested by another test known as log-rank test and Cox proportion hazard test. In log-rank test we calculate the expected number of events in

each group. This is how we will move forward in our analysis next week when we begin to build our models.

Contributions

For this paper Sarah wrote the introduction, variable exploration section, and created/analyzed the Kaplan-Meier survival curves. Nathaniel worked on the interpretation of the Kaplan-Meier curves, and finalized the plots with details such as axis labels. Inhee worked on data visualizations, variable exploration, and exploratory data analysis, and edited the entire report.

References

http://influentialpoints.com/Training/coxs_proportional_hazards_regression_model-principles-properties-assumptions.htm#modmch

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/

http://dwoll.de/rexrepos/posts/survivalKM.html