In this dataset, the complete career histories of RAS staff were obtained. These data include 242 unique individuals (defined by *uid*) and 272 unique individual-level employment spells (defined by *rasid*). The expanded data file consists of 1,256 cases, 49 voluntary quits, and eight involuntary quits.

The first question assignment requires data analysis using graphical displays (Kaplan-Meier survival plots) and descriptive statistics that can document how employee turnover varies with individual variables.

Preliminary Data Manipulation

As seen in **(Figure A)**, all the variables were scanned and familiarized before embarking on any analysis task. Based on the variables' description and nature, each variable's measure is changed to scale, nominal, or ordinal, as shown in **(Figure B)**. Based on the instruction, a natural logarithm is applied to the variable *annualrt* to ensure that the data distribution resembles a less skewed shape and named as *annualrt_LN*, as shown in **(Figure C)**. Both **(Figure D)** and **(Figure E)** shows the graphical representation of *the AnnualRt* variable in terms of histograms. The histogram of the variable *annualrt_LN* resembles a more normal shape than the original variable. As instructed, the variable *Rank* is also recoded into a different variable by classifying the *rank IV* and above into the same level as shown in **(Figure F)** and **(Figure G)**.

Descriptive Statistics

Looking at each variable's descriptive statistics in the dataset is also a critical preliminary step before embarking on any fundamental analysis. Descriptive statistics for each variable in the data are computed as shown in **(Figure 1)**. It's interesting to notice that *Annual Rt* has a standard deviation value of 17,717 while *annualrt_LN* (natural logarithm applied version) has a standard deviation value of only 0.24. The frequency distribution of the variable *Rank* (recoded version) is computed on **(Figure 2)**. It is seen that frequency of *Rank I* and *Rank IV* (and above) is quite comparable, while *Rank II* and *Rank III* account for a majority of the data.

In *Staff works in pre- or post-award stage* variable, it is seen that 'other' level accounts for little more than 10 percent of the data, while 'Pre' and 'Post' account for little less than 90 percent of the data as shown in the **(Figure 3)**. In the *Division* variable, as shown in **(Figure 4)**, it is observed that most data points come from either School of Medicine and School of Public Health, accounting for more than 75 percent of the data. In the *Department* variable, as shown in **(Figure 5)**, it is observed that most data points come from either the School of Medicine or School of Public Health, both of these departments accounting for more than 75 percent of the data.

In the *Gender* variable, as shown in **(Figure 6)**, it is observed that most data points come from *female*s, accounting for more than 80 percent of the data. In the *Supervisor Gender* variable, as shown in **(Figure 7)**, it is observed that most data points come from class 1 *(Female Supervisor)*, accounting for more than 85 percent of the data. Both **(Figure 6)** and **(Figure 7)** suggest that females make up most of the data points both in terms of survey participants and participants' supervisors. In the *RAS* variable, as shown in **(Figure 8)**, it is observed that most data points come from the department within the School of Medicine, accounting for more than 50 percent of the data. In the *Origin of RAS employee* variable, as shown in **(Figure 9)**, it is observed that most data points come from the same division, indicating that more than 50 percent of the participants come from the same division.

In the *White* variable, as shown in **(Figure 10)**, it is observed that data is equally distributed, meaning there is no significant difference in the number of white versus non-white survey participants. However, when it comes to supervisor race, as shown in the *Supervisor Race* variable shown in **(Figure 11)**, it is observed that most supervisors are white, accounting for almost 75 percent of the data. In the *Education Level* variable, as shown in **(Figure 12)**, it is observed that most participants hold either a Bachelor's degree or Master's degree, both of these departments accounting for more than 80 percent of the data.

In the *Treatment Group* variable, as shown in **(Figure 13)**, it is observed that most data points come from the non-treated group, accounting for more than 70 percent of the data. In the *f-quit* variable, as shown in **(Figure 14)**, it is observed that most data points come from the participants who have not quit (or did not experience turnover), accounting for more than 95 percent of the data.

Kaplan-Meier Plots (with Log-rank test):

As shown in **(Figure 15)**, Kaplan-Meier with the log-rank test is performed for all types of turnover (both 1 and 2), including voluntary and involuntary turnover. The variable *t_quit*, duration at the end of the spell, is chosen as the time variable and *f_quit* as the status variable.

Individual variable: Rank

Descriptive statistics are computed as shown in **(Figure 16)** and **(Figure 17)** which document how employee turnover varies with the individual variable *Rank*. To briefly articulate the log-rank test's purpose, it is considered a popular method to test the null hypothesis of no difference in survival outcomes between two or more independent samples, comparing the entire survival experience among groups by comparing whether the survival curves are identical (overlapping) or not. A log-rank test was computed to determine if there were differences in the survival distribution for the different types of intervention: *Rank I*, *Rank II*, *Rank III*, and *Managerial*. The survival distributions for the four groups were statistically significantly different ($p < 0.001$), as shown in **(Figure 18)**. Based on the survival curve shown in **(Figure 19)**, there is a systematic difference across this attribute. We can see from our plot that the cumulative survival proportion appears to be much higher in the *Rank III* and *Mgr* group compared to the *Rank II* and *Rank I*, which appear to differ considerably. It would appear that having a higher Rank significantly prolongs the time until participants quit (i.e., the event) compared to the relatively lower ranks. This survival curve is considered statistically significant based on the significance level of the log-rank test ($p < 0.001$).

Individual variable: Staff Works in pre- or post-award stage

Descriptive statistics are computed as shown in **(Figure 20)** and **(Figure 21)**, which document how employee turnover varies with the individual variable *Staff Works in pre- or post-award stage*. A log-rank test was computed to determine if there were differences in the survival distribution for the different types of stage: *Pre, Post,* and *Other*. The three groups' survival distributions were statistically significantly different ($p = 0.338$), as shown in **(Figure 22)**. Based on this result, the survival curve is not shown since the log-rank test indicates that the variation in employee turnover is not significant with this individual variable.

Individual variable: Age (binned)

Descriptive statistics are computed as shown in **(Figure 23)** and **(Figure 24)**, which document how employee turnover varies with the individual variable *Age (binned)*. When a log-rank test was computed to determine if there were differences in the survival distribution for the different types of intervention: *1 (20 to 30 y/o), 2 (30 to 40 y/o), 3 (40 to 50 y/o), 4 (50 to 60 y/o),* and *5 (60 y/o and above)*. The survival distributions for the five groups are statistically significantly different, as indicated by the significance level ($p < 0.001$) shown in **(Figure 25)**. Based on the survival curve shown in **(Figure 26)**, there is a systematic difference across this attribute. We can see from our plot that the cumulative survival proportion appears to be much higher in the *Age 4.00* and *Age 5.00* group compared to the *Age 1.00*, *Age 2.00*, *and Age 3.00*, which appear to differ considerably. It would appear that having older ages significantly prolong the time until participants quit (i.e., the event) compared to the relatively younger generations. This survival curve is considered statistically significant based on the significance level of the log-rank test ($p < 0.001$).

Individual variable: Gender

Descriptive statistics are computed as shown in **(Figure 27)** and **(Figure 28)**, which document how employee turnover varies with the individual variable *Gender*. A log-rank test was computed to determine if there were differences in the survival distribution for the different types of variables: *Male* and *Female*. The survival distributions for the two groups are not shown to be statistically different ($p = 0.799$), as shown in **(Figure 29)**. Based on this result, the survival curve is not shown since the log-rank test indicates that the variation in employee turnover is not significant with this individual variable.

Individual variable: Supervisor Gender

Descriptive statistics are computed as shown in **(Figure 30)** and **(Figure 31)** that document how employee turnover varies with the individual variable *Supervisor Gender*. A log-rank test was computed to determine if there were differences in the survival distribution for the different types of supervisor gender: *Male Supervisor* and *Female Supervisor*. The survival distributions for the two groups are shown to be statistically different ($p = 0.026$), as shown in **(Figure 32)**. Based on the survival curve shown in

(Figure 33), there is a systematic difference across this attribute. From our plot, we can see that the cumulative survival proportion appears to be much higher in the *male supervisor* group than the *female supervisor* group, which seems to differ considerably. It would appear that having a male supervisor significantly prolong the time until participants quit (i.e., the event) compared to having a *female supervisor*. Again, based on the log-rank test's significance level, this survival curve is considered statistically significant.

Individual variable: RAS Type

Descriptive statistics are computed as shown in **(Figure 34)** that document how employee turnover varies with the individual variable *RAS Type*. A log-rank test was calculated to determine if there were differences in the survival distribution for the ten different RAS departments. The ten groups' survival distributions are not shown to be statistically different ($p = 0.054$), as shown in **(Figure 35)**. Based on this result, the survival curve is not shown since the log-rank test indicates that the variation in employee turnover is not significant with this individual variable.

Individual variable: Origin of RAS Employee

Descriptive statistics are computed as shown in **(Figure 36)** and **(Figure 37)**, which document how employee turnover varies with the individual variable *Origin of RAS Employee*. A log-rank test was computed to determine if there were differences in the survival distribution for RAS employees' different origins: *New entry*, *Different division,* and *Same division*. The survival distributions for the three groups are shown to be statistically different ($p < 0.001$), as shown in **(Figure 38)**. Based on the survival curve shown in **(Figure 39)**, there is a systematic difference across this attribute. We can see from our plot that the cumulative survival proportion appears to be much higher in the *Same division* group compared to the *New entry* and *Different division* group, which appear to differ considerably. It would appear that coming from the same division significantly prolongs the time until participants quit (i.e., the event) compared to coming from a different division or coming as a new entry. Again, based on the log-rank test's significance level, this survival curve is considered statistically significant.

Individual variable: Tenure (binned)

The tenure variable is originally continuous. This variable is converted into a categorical variable to perform survival analysis by binning accordingly, as shown in **(Figure 40)**. Then, descriptive statistics are computed as shown in **(Figure 41)** and **(Figure 42)** which document how employee turnover varies with the individual variable *TenureTenure (binned)*. A log-rank test was computed to determine if there were differences in the survival distribution for the different tenure categories from *Level 1.00* to *Level 7.00*. The seven groups' survival distributions are shown to be statistically different ($p = 0.019$), as shown in **(Figure 43)**. Based on the survival curve shown in **(Figure 44)**, there is a systematic difference across this attribute. We can see from our plot that the cumulative survival proportion appears to be much higher in the *Level II, III, IV, and V* group compared to the *Level I, VI,* and *VII* group, which appear to differ considerably. It would appear that coming from Tenure that ranges from 51 to 250 significantly prolongs the time until participants quit (i.e., the event) compared to having TenureTenure that ranges from either 0 to 50 or 251 to the highest TenureTenure. Again, based on the log-rank test's significance level, this survival curve is considered statistically significant.

Individual variable: White

Descriptive statistics are computed as shown in **(Figure 45)** and **(Figure 46A)** that document how employee turnover varies with the individual variable *White*. A log-rank test was calculated to determine if there were differences in the survival distribution for the different races: *Non-white* and *White*. The survival distributions for the two groups are shown to be statistically different ($p = 0.023$), as shown in **(Figure 46B)**. Based on the survival curve shown in **(Figure 47)**, there is a systematic difference across this attribute. From our plot, we can see that the cumulative survival proportion appears to be much higher in the *white* group than the *non-white* group, which appears to differ considerably. It would appear that being a white person significantly prolongs the time until participants quit (i.e., the event) compared to being a non-white person. Again, based on the log-rank test's significance level, this survival curve is considered statistically significant.

Individual variable: Supervisor Race

Descriptive statistics are computed as shown in **(Figure 48)** and **(Figure 49)**, which document how employee turnover varies with the individual variable *Supervisor Race*. A log-rank test was computed to determine if there were differences in the survival distribution for the two different supervisor races: *non-white supervisor* and *white supervisor*. The survival distributions for the

two groups are not shown to be statistically different (p = 0.133), as shown in **(Figure 50)**. Based on this result, the survival curve is not shown since the log-rank test indicates that the variation in employee turnover is not significant with this individual variable.

Individual variable: Education Level

Descriptive statistics are computed as shown in **(Figure 51)** and **(Figure 52)** which document how employee turnover varies with the individual variable *Education Level*. A log-rank test was computed to determine if there were differences in the survival distribution for the four different education levels: *HS and associate's degree, Bachelor's degree, Master's degree,* and *Doctoral degree*. The four groups' survival distributions are not shown to be statistically different (p = 0.120), as shown in **(Figure 53)**. Based on this result, the survival curve is not shown since the log-rank test indicates that the variation in employee turnover is not significant with this individual variable.

Individual variable: Treatment Group

Descriptive statistics are computed, as shown in **(Figure 54)** and **(Figure 55),** that document how employee turnover varies with the individual variable *Treatment Group*. A log-rank test was computed to determine if there were differences in the survival distribution for two groups *0 (not treated)* and *1 (treated)*. The survival distributions for the two groups are not shown to be statistically different (p = 0.439), as shown in **(Figure 56)**. Based on this result, the survival curve is not shown since the log-rank test indicates that the variation in employee turnover is not significant with this individual variable.

In the following question, statistical models are used to identify factors that explain variations in employee turnover rate using a Cox regression model) for *a. All types of quits*, *b. Only voluntary quits*, and *c. Only involuntary quits*.

<u>All type of quits:</u>

A model with all variables except for Rank and Tenure:

Notice that *annualrt_LN*, natural logarithm form of *annualrt*, is entered into a Cox regression. *Anualrt* is also highly correlated with *Rank*, and so is *age* with TenureTenure. Only one of the variables is included (in this case, only *annualrt* and *age*) while specifying categorical variables when performing cox regression, as shown in **(Figure 57)**. In **(Figure 58)**, we can see that out of 1,256 total events, 1,167 events are censored, and 33 cases are dropped. One of the outputs from running cox regression includes the Omnibus Tests of Model Coefficients, which is used to check that the new model (with explanatory variables included) is an improvement over the baseline model. It uses chi-square tests to see if there is a significant difference between the Log-likelihoods (specifically the -2LLs) of the baseline model and the new model. Based on the significance level (p < 0.001), it is safe to regard this new model to be an improvement over the baseline model, as shown in **(Figure 59)**. When interpreting the coefficients of this model, as shown in **(Figure 60)**, we can see that only a few variables such as *Supervisor Gender*, *RAS unit (3)*, *Origin of RAS employee (1)*, *Origin of RAS employee (2)*, *Supervisor race*, *annualrt_LN* are statistically significant. Based on interpreting the Exponentiated beta value (odds ratio) of each variable with statically substantial values, it is evident that having a female supervisor will multiply the relative risk of turnover by 0.272 (or 27.2 percent). The RAS unit (3), or SOM Medicine RAS, will multiply the relative risk by 0.134 (or 13.4 percent). The origin of RAS employees seems to matter a lot: coming from different divisions, multiply the relative risk by 4.438 (or 443.8 percent), and starting as a new RAS entry will multiply the risk by 2.954 (or 295.4 percent). Additionally, having a white supervisor will multiply the relative risk of the turnover by 2.548 (or 254.8 percent). Lastly, one unit increase of annual pay rate (with natural logarithm applied) will multiply the relative risk of turnover by 17.5 percent. The covariates mean of each variable is also computed, as shown in **(Figure 61)**.

With Only Rank and Tenure in a Model:

Cox regression is re-calculated, but this time with only the variable *Rank* and *Tenure*. Based on the Omnibus test of model coefficients as shown in **(Figure 62)**, a new model is considered an improvement over the baseline model. Based on interpreting the Exponentiated beta value (odds ratio) of each variable with statically significant values as shown in **(Figure 63)**, it is evident that having a Rank (1), which is equivalent to Rank II, will multiply the relative risk of the turnover (compared to the base outcome) by 6.866 (or 686.6 percent). Similarly, one month increase of Tenure at Emory will multiply the relative risk of turnover (compared to the base outcome) by 0.996 (or 99.6 percent).

Only voluntary quits (1):

As shown in **(Figure 64)**, cox regression is computed again but this time, only specifying the Status variable *f_quit* to be 1 (or include only voluntary turnover).

A model with all variables except for Rank and Tenure:

Case processing summary is shown in **(Figure 65)**, from which we can infer that 1,174 out of 1,256 total events are censored while 33 events are dropped. The coding of each categorical output is shown in **(Figure 66)**. Based on the significance level ($p < 0.001$) indicated on the Omnibus Tests of Model Coefficients shown in **(Figure 67)**, it is safe to regard this new model to be an improvement over the baseline model. When interpreting the coefficients of this model, as shown in **(Figure 68)**, we can see that only a few variables such as *supervisor gender RAS unit (3), Origin of RAS employee, Supervisor Race,* and *Annualrt_LN* are statistically significant. Based on interpreting the Exponentiated beta value (odds ratio) of these five variables, it is evident that having a female supervisor will multiply the relative risk of the turnover (compared to the base outcome) by 0.183 (or 18.3 percent). Similarly, the SOM Medicine RAS unit will multiply the relative risk of the turnover by 0.151 (or 15.1 percent). RAS employees' origin seems to matter a lot: coming from different divisions will multiply the relative risk of the turnover by 4.121 (or 412.1 percent), and coming from new entry will multiply the risk of turnover by 2.855 (or 285.5 percent). Having a white supervisor will also multiply the relative risk of the turnover by 2.642 (or 264.2 percent), and lastly, having one unit increase of annual pay rate (with natural logarithm applied) will multiply the relative risk of turnover by 0.217 (or 21.7 percent). Covariate means of these variables are shown in **(Figure 69)**.

A model with only Rank and Tenure:

Case processing summary is shown in **(Figure 70)**, from which we can infer that 1,192 out of 1,256 total events are censored while 15 events are dropped. The output of the *rank* variable coded into a categorical variable is shown in **(Figure 71)**. Based on the significance level ($p < 0.001$) indicated on the Omnibus Tests of Model Coefficients shown in **(Figure 72)**, it is safe to regard this new model to be an improvement over the baseline model. When interpreting this model's coefficients, as shown in **(Figure 73)**, we can see that only a few variables such as *Rank (1)* and *Tenure at Emory* are statistically significant. Based on interpreting the Exponentiated beta value (odds ratio) of these two variables, it is evident that having a Rank (I), which is equivalent to Rank II, will multiply the relative risk of the turnover (compared to the base outcome) by 5.784 (or 578.4 percent). Similarly, one month increase of Tenure at Emory will multiply the relative risk of the turnover (compared to the base outcome) by 0.995 (or 99.5 percent). Covariate means of these two variables are shown in **(Figure 74)**.

Only involuntary quits (2):

A model with all variables except for Rank and Tenure:

Lastly, a cox regression is computed but this time only specifying the Status variable *f_quit* to be 2 (or include only involuntary turnover). The case processing summary shown in **(Figure 75)** indicates that 1,188 out of 1,256 total events are available in the analysis, and 68 cases are dropped. Based on the significance level ($p = 0.013$) indicated on the Omnibus Tests of Model Coefficients shown in **(Figure 76)**, it is safe to regard this new model to improve the baseline model. When interpreting this model's coefficients, as shown in **(Figure 77)**, we can see that only a few variables such as *Gender, Origin of RAS employees, and White* are statistically significant. Based on interpreting the Exponentiated beta value (odds ratio) of these three variables, it is evident that being a female will increase the relative risk of the turnover (compared to the base outcome) by 11.995 (or 1199.5 percent). RAS employee's origin seems to matter a lot: coming from the different divisions will multiply the relative risk of the turnover by 21.546 (or 2154.6 percent). Lastly, being a white person will multiply the relative risk by 26.542 (or 2654.2 percent). Covariate means of these variables are shown in **(Figure 78)**.

A model with only Rank and Tenure:

Based on the significance level ($p = 0.073$) indicated on the Omnibus Tests of Model Coefficients shown in **(Figure 79)**, it is not safe to assume this new model to be an improvement over the baseline model. When interpreting this model's coefficients, as shown in **(Figure 80)**, we can see that no variables are considered to be statistically significant; therefore, further analysis will not be made.