RAS Staff at Emory University

In this assignment, we are working with the complete career histories of RAS staff at Emory University. In *assign2_ratings.sav*, performance evaluations are included. In *assign2_promotion_merit.sav*, information on the number of merit awards received by each employee and the occurrence and time to their first promotion is included. It is important to note that the types of dependent variable and independent variables when performing the statistical test. First , we can determine the type of statistical test based on the types of dependent variable. When the dependent variable is continuous (e.g., sales), we can simply use OLS regression. When the dependent variable is ordinal (e.g., customer loyalty ratings), we can use ordered logit and probit regression. When dependent variables is in forms of count variables (e.g., number of items scanned, number of merit awards), it is important to check the over-dispersion; Then, we can determine whether to use either Poisson regression (when mean is roughly equal to variance) and negative binomial (when mean is quite divergent from the value of the variance). Even though it may not be discussed in detail given the scope of this course, we can use zero-inflated models when using dependent count variables where the instances of zero is so much bigger than other instances (e.g., fishing example: number of fish caught by people in the park – 0 are largely determined by population of people who didn't go to park). Types of independent variables also matter, and if necessary these should be recoded into categorical variables. In addition to regression tests, we should also consider the types of both dependent and independent variables when performing univariate statistical analysis. When dependent variable is categorical, we can perform chi-square test. When dependent variable is continuous and independent variable is binary, we can perform independent sample t-test. And when dependent variable is continuous and independent variable has more than two categories, we can perform ANOVA. To summarize the types of statistical test to consider, we can consider chi-square test (when we perform univariate test on count variables) and Poisson/Negative binomial regression (when we perform multivariate analysis on count variables). On the other hand, we can consider log-rank test (when we perform univariate test on time variables) and survival analysis: Cox regression (when we perform multiple regression on time variables). It is always important to keep in mind that we should first correctly specify the independent variables in the multiple regression (whether we are going to use continuous or categorical independent variables and to gauge if categorization of continuous variable into discrete variable is reasonable).

First, we are asked to display graphical representation and descriptive statistics to document how the review ratings vary with individual variables**.** It is important to first look at the variables to understand the type and nature of the variables before performing any statistical analysis. In the **(figure 1)**, we can see the review year is categorical variable, but distribution of the data is highly skewed, meaning most data is highly skewed on the late 2010's. Therefore, it wouldn't be a bad idea to re-categorize this variable to ensure that we have equal representation of each data point as shown above. Looking at **(Figure 2)**, we can see that the variable *'Department'* and *'RAS unit'* basically contains same set of information but in a different manner (more concise form of variable names in the case of RAS unit). Therefore, it is important to note that we are only including one of the two variables, but not two, when performing statistical analysis including regression. In the **(Figure 3)**, we can see that across all departments in the data only two departments within SOM have the presence of the treatment group (namely, *SOM: Basic Science RAS* and *SOM: Medicine RAS*). Thus, it is important to keep this in mind when performing the analysis. In the **(Figure 4)**, we can see that there are a total of 5 divisions across this data. When cross tabulated with all departments, it is noticeable that School of Medicine is the only division in which there are multiple departments within. In the **(Figure 5)**, we see the descriptive statistics for each variable with exception of *uid* and continuous variables. From the figure, we can see that the variables such as *'Gender'*, *'White'*, *'Supervisor gender'*, *'Supervisor race'*, and *'Treatment group'* have only 0 and 1 for their values, indicating that they are binary variables. Whereas, variables such as *'Education level'*, *'RAS unit'*, *'Rank'*, *'Staff works…'*, *'Review rating'* and *'Origin of RAS employee'* are still categorical variables but have more than two categories. When looking at the distribution of *'Gender'* column **(Figure 6)**, it is noticeable that there are about four times as many more females (487; 82.5 percent) than males (103; 17.5 percent) in this dataset. When looking at the distribution of *'White'* column **(Figure 7)**, it is noticeable that there are roughly similar representation of non-white (311) versus white (279) participants in the dataset. When looking at the distribution of *'Education level'* column **(Figure 8)**, it is noticeable that a majority of the participants have either bachelor's or master's degree (85.6 percent); whereas, HS and Associate's degree make up only 11.5 percent, and doctoral degree make up only 1.4 percent. When looking at the distribution of *'Supervisor gender'* column **(Figure 9)**, it is noticeable that there are more than six times as many more females (507; 85.9 percent) than males (83; 14.1 percent) in this dataset. Conversely, when looking at the distribution of *'Supervisor race'* column **(Figure 10)**, it is noticeable that there are more than three times as many more white supervisors (430; 72.9 percent) than non-white supervisors (160; 27.1 percent) in this dataset. When looking at the distribution of *'Division'* column **(Figure 11)**, it is noticeable that a majority of the participants is affiliated with either SOM (65.4 percent) or SOPH (19.2 Percent). While, the rest of the division such as Emory College (5.8 percent), Research Administration (2.0 percent), Yerkes National Primate Research Center (7.6 percent) only make up the small percentages of the total division. When looking at the distribution of *'Title'* column **(Figure 12)**, it is noticeable that a majority of the participants is employed as Research Admin (Post-Award II & III and Pre-Award II & III), making up more than 50 percent of the whole title composition. When looking at the distribution of *'Rank'* column **(Figure 13)**, it is noticeable that a majority of the participants has rank of either II (32.9 percent) or III (41.9 percent), making up more than 70 percent when combined together. It is important to notice that the rank Sr. Dir is the highest rank within this data. When looking at the distribution of *'Staff works in pre/post-award stage'* column **(Figure 14)**, it is noticeable that data is fairly distributed between 0:Pre (37.5 percent) and 1:Post (50.2 percent), making more than 85 percent of the whole when combined. When looking at the distribution of *'Review rating'* column **(Figure 15)**, it is noticeable that two responses - 2: Meets expectations (32 percent) and 4: Exceeds expectation (53.1 percent) – make up more than 80 percent of the data when combined. When looking at the distribution of *'Origin of RAS Employee'* column **(Figure 16)**, it is noticeable that the most employees are from the same division (62.5 percent), followed by different division (20.7 percent) and others (16.8 percent). When looking at the distribution of *'Treatment Group'* column **(Figure 17)**, it is noticeable that the most employees are not treated (447; 75.8 percent) compared to the percentages of those who are treated (143; 24.2 percent). The column *'age'* is originally encoded as continuous variable. And for the purpose of further analysis, it is recoded to a categorical variable to new column named *'age_binned'* as shown in **(Figure 18)**. When looking at the distribution of *'Age Binned'* column **(Figure 19)**, it is noticeable that the most employees are between class 3 (40 through 49.99 y/o; 30.7 percent) and class 4 (50 through 59.99 y/o; 30.3 percent). The column *'Tenure at Emory'* is originally encoded as continuous variable. And for the purpose of further analysis, it is recoded to a categorical variable to new column named *'tenure_binned'* as shown in the **(Figure 20)**. When looking at the distribution of *'Tenure Binned'* column shown in the **(Figure 21)**, it is noticeable that the most employees are in the class 6 (250 through 299.99; 59.2 percent) and class 1 (0 through 49.99; 29.3 percent), making up more than 90 percent of the distribution when combined. Finally, when looking at the distribution of *'Review Year Binned'* column shown in the **(Figure 22)**, it is noticeable that each class is roughly equally distributed. Now that we have sufficiently explored the data, cross tabulation is the next step where the chi-square test is performed. It is important to note that we need to input both independent and dependent *(review ratings)* variable as categorical variables. Individual variables are simply the variables that can uniquely identify each individual. Referring to the variable code book, variables including *age_binned, gender_binned, white, educlv (education level), supvgender (supervisor gender), supvwhite (supervisor race), rastype, rank, award, ras_entrymode*, *ras_treat*, and *review_year_binned* are independent variables that can uniquely describe each individual and that can have predictive power on the dependent variable *(review ratings)*. The variables *age*, *gender, review_year* were initially coded as the continuous variable, and hence they are recoded into into categorical variable by

binning based on their respective distribution. As seen in the **(Figure 23)**, cross-tabulation with chi-square is computed for each categorical variable against 'review rating'.

Result:

The first table is the Case Processing summary as shown in the **(Figure 24)**, it tells us the number of valid cases used for analysis. There exist a few missing values when *'Review rating'* is cross tabulated with *'Education level'*. Only cases with non-missing values for both *'Review rating'* and *'Education level'* can be used in the test.

Review rating * Gender: This table shown in the **(Figure 25)** the crosstabulation and chi-square test results between review ratings and gender. It is interesting to notice that there are more female responses (82.5 percent) of review ratings than male responses (17.5 percent). Based on the chi-square test as shown in the **(Figure 26)**, the value of the test statistic is 1.157. The corresponding p-value of the test statistic is $p = 0.763$. Since the p-value is greater than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest an association between gender and review ratings. Based on the results, we can state the following: No association was found between gender and review ratings. This relationship can also be graphically represented as shown in the **(Figure 27)**.

Review rating * White: The cross tabulation shown in the **(Figure 28)** displays chi-square test results between review ratings and white (race). It is noticed that there is roughly equal amount of representation of review ratings between non-white (311; 52.7 percent) and white (279; 47.3 percent). Based on the chi-square test as shown in the **(Figure 29)**, the value of the test statistic is 17.809. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is less than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between gender and white. Based on the results, we can state the following: association was found between gender and race (white or not). This relationship can also be graphically represented as shown in the **(Figure 30)**.

Review rating * Education level: The **(Figure 31)** shows the crosstabulation and chi-square test results between review ratings and education level. It is interesting to notice that a majority of participants hold bachelor's degree, out of which meets (34.0 percent), exceeds (52.3 percent) or far exceeds the expectation (12.3 percent). Based on the chi-square test as shown in the **(Figure 32)**, the value of the test statistic is 9.470. The corresponding p-value of the test statistic is $p = 0.395$. Since the p-value is greater than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest an association between education level and review ratings. Based on the results, we can state the following: No association was found between gender and education level. This relationship can also be graphically represented as shown in the **(Figure 33)**.

Review rating * Supervisor gender: The **(Figure 34)** shows the crosstabulation and chi-square test results between review ratings and supervisor gender. It is interesting to notice that a majority of the supervisor gender is female and that most female supervisor exceeds expectation (52.9 percent) of review ratings. Based on the chi-square test as shown in the **(Figure 35)**, the value of the test statistic is 7.266. The corresponding p-value of the test statistic is $p = 0.064$. Since the p-value is greater than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest an association between supervisor gender and review ratings. Therefore, based on the results, we can state the following: No association was found between supervisor gender and review ratings. This relationship can also be graphically represented as shown in the **(Figure 36)**.

Review rating * Supervisor race: The table shown in the **(Figure 37)** displays the crosstabulation and chi-square test results between review ratings and supervisor race. It is interesting to notice that a majority of the supervisor race is white and that most white supervisor exceeds expectation (54.9 percent) of review ratings. Based on the chi-square test as shown in the **(Figure 38)**, the value of the test statistic is 11.926. The corresponding p-value of the test statistic is $p = 0.008$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between supervisor race and review ratings. Therefore, based on the results, we can state the following: association was found between supervisor race and review ratings. This relationship can also be graphically represented as shown in the **(Figure 39)**.

Review rating * Division: The **(Figure 40)** above shows the crosstabulation and chi-square test results between review ratings and division. It is interesting to notice that a majority of the review rating is conducted in the school of medicine (SOM) and that most participants within SOM exceeds expectation (58.4 percent) of review ratings. Based on the chi-square test as shown in the **(Figure 41)**, the value of the test statistic is 14.061. The corresponding p-value of the test statistic is $p = 0.297$. Since the p-value is bigger than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest an association between division and review ratings. Therefore, based on the results, we can state the following: No association was found between division and review ratings. This relationship can also be graphically represented as shown in the **(Figure 42)**.

Review rating * Department: The **(Figure 43)** shows the crosstabulation and chi-square test results between review ratings and department. It is interesting to notice that a majority of the participants is affiliated with the department within SOM: Cancer RAS, Medicine RAS, Pediatric RAS and SPH: Research Admin. Most of them within these department either meets expectation or exceed expectation of review ratings. Based on the chi-square test as shown in the **(Figure 44)**, the value of the test statistic is 49.641. The corresponding p-value of the test statistic is $p = 0.005$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. We can conclude that there may be enough evidence to suggest an association between department and review ratings. Therefore, based on the results, we can state the following: association is found between department and review ratings. This relationship can also be graphically represented as shown in the **(Figure 45)**.

Review rating * RAS unit: The **(Figure 46)** shows the crosstabulation and chi-square test results between review ratings and RAS unit. It is interesting to notice that the results look almost identical to the previous cross-tabulation analysis between review ratings and department. Since these two columns contain similar information, it is important to keep in mind that we only use one or the another when performing regression to predict a particular dependent variable. Based on the chi-square test as shown in the **(Figure 47)**, the value of the test statistic is 49.641. The corresponding p-value of the test statistic is $p = 0.005$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between RAS unit and review ratings. Therefore, based on the results, we can state the following: No association was found between RAS unit and review ratings. Of course, the result looks almost identical to that of department. This relationship can also be graphically represented as shown in the **(Figure**

48).

Review rating * Title: The **(Figure 49)** shows the crosstabulation and chi-square test results between review ratings and title. Notice that there are multiple levels within title that it is hard to interpret the result. It is interesting to notice that a majority of the title falls under Research Admin Pre/Post-Award and that most of them falls under meeting expectation of the review ratings. Based on the chi-square test as shown in the **(Figure 50)**, the value of the test statistic is 169.099. The corresponding p-value of the test statistic is $p = 0.002$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between title and review ratings. Therefore, based on the results, we can state the following: association was found between title and review ratings. This relationship can also be graphically represented as shown in the **(Figure 51)**.

Review rating * Rank (increases from I to Sr Dir): The table shown in the **(Figure 52)** shows the crosstabulation and chi-square test results between review ratings and rank. It is interesting to notice that a majority of the rank falls under either II (32.9 percent) or III (41.9 percent) and most of participants under these ranks either meet or exceed expectations of review ratings. Based on the chi-square test as shown in the **(Figure 53)**, the value of the test statistic is 57.303. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between rank and review ratings. Therefore, based on the results, we can state the following: association was found between rank and review ratings. This relationship can also be graphically represented as shown in the **(Figure 54)**.

Review rating * Staff works in pre- or post-award stage: The table shown in the **(Figure 55)** shows the crosstabulation and chi-square test results between review ratings and Staff works in pre- or post-award stage. It is interesting to notice that a majority of the staffs works in a post-stage and that most of them meet expectation (57.1 percent) of review ratings. Based on the chi-square test as shown in the **(Figure 56)**, the value of the test statistic is 27.799. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between staff works in pre/post-award and review ratings. Therefore, based on the results, we can state the following: strong association was found between staff works in pre/post-award and review ratings. This relationship can also be graphically represented as shown in the **(Figure 57)**.

Review rating * Origin of RAS employee: The **(Figure 58)** shows the crosstabulation and chi-square test results between review ratings and origin of RAS employee. It is interesting to notice that a majority of the staffs is from same vision and that most of them exceed expectation (55.3 percent) of review ratings. Based on the chi-square test as shown in the **(Figure 59)**, the value of the test statistic is 23.986. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between origin of RAS employee and review ratings. Therefore, based on the results, we can state the following: strong association was found between origin of RAS employee and review ratings. This relationship can also be graphically represented as shown in the **(Figure 60)**.

Review rating * Treatment group: The table shown in the **(Figure 61)** shows the crosstabulation and chi-square test results between review ratings and treatment group. It is interesting to notice that a majority of the staffs is not administered with treatment (75.8 percent) and that most of them exceed expectation (55.0 percent) of review ratings. Based on the chi-square test as shown in the **(Figure 62)**, the value of the test statistic is 5.388. The corresponding p-value of the test statistic is $p < 0.145$. Since the p-value is bigger than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest an association between treatment groups and review ratings. Therefore, based on the results, we can state the following: no association was found between treatment groups and review ratings. This relationship can also be graphically represented as shown in the **(Figure 63)**.

Review rating * Age Binned: The **(Figure 64)** shows the crosstabulation and chi-square test results between review ratings and age. It is interesting to notice that a majority of the staffs falls under age group 3 (40-49.99 y/o) and 4 (50-59.99 y/o) and that most of them exceed expectation (more than 50 percent) of review ratings. Based on the chi-square test as shown in the **(Figure 65)**, the value of the test statistic is 23.931. The corresponding p-value of the test statistic is $p = 0.066$. Since the p-value is bigger than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest an association between age and review ratings. Therefore, based on the results, we can state the following: no association was found between age and review ratings. This relationship can also be graphically represented as shown in the **(Figure 66)**.

Review rating * Tenure Binned: The table shown in the **(Figure 67)** shows the crosstabulation and chi-square test results between review ratings and tenure. It is interesting to notice that a majority of the staffs falls under bin 1 (0-49.99 tenure) and bin 6 (300 and up tenure) and that most of them exceed expectation of review ratings. Based on the chi-square test as shown in the **(Figure 68)**, the value of the test statistic is 18.453. The corresponding p-value of the test statistic is $p = 0.240$. Since the p-value is bigger than our chosen significance level ($\alpha = 0.05$), we do not reject the null hypothesis. Rather, we conclude that there is not enough evidence to suggest an association between tenure status and review ratings. Therefore, based on the results, we can state the following: no association was found between tenure status and review ratings. This relationship can also be graphically represented as shown in the **(Figure 69)**.

Review rating * Review Year Binned: As shown in the **(Figure 70)**, it is notable that most review ratings fall under the bin 4 and 5 and that most of the review ratings conducted during these years show most review ratings exceeding expectations. Based on the chi-square test as shown in the **(Figure 71)**, the value of the test statistic is 29.957. The corresponding p-value of the test statistic is $p = 0.003$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between review years and review ratings. Therefore, based on the results, we can state the following: strong association was found between review years and review ratings. This relationship can also be graphically represented as shown in the **(Figure 72)**.

In the next question, we are asked to use statistical models to identify factors that explain variations in the annual review ratings given to employees. It is important to first determine the type of dependent variables we are exploring in order to answer this question. It is noted that the dependent variable *'annual review ratings'* is considered to be ordinal variable since it is a categorical variable for which the possible values are ordered. Therefore, we need to consider the ordered logit model (also known as ordinal regression model) as shown in the **(Figure 73)**, which is a regression model for ordinal dependent variables. When running ordinal logistic regression, we need to ensure that data "passes" three assumptions that are required to give us a valid result.

- Assumption #1: Dependent variable should be measured at the ordinal level.
- Assumption #2: One or more independent variables should be continuous, ordinal or categorical (including dichotomous variables). Conversely, ordinal independent variables must be treated as being either continuous or categorical, meaning they cannot be treated as ordinal variables.
- Assumption #3: There should be no multicollinearity. Testing multicollinearity may require creating dummy variables.

In addition to "passing" these assumptions, it is important to use the variables that are deemed significant from the chi-square test performed in the previous question. In this case, the variables that are considered significant and should be entered into the ordinal logistic regression model are: *'reviewyear_binned'*, *'ras_entrymode'*, *'award'*, *'rank'*, *'title'*, *'department'*, *'supvwhite' and 'white'*. However, notice that *'department'* and *'title'* are textual variables and should be replaced by *'rastype'*, which contains basically the same information about the RAS unit. Also, the variable 'rank' has the ordered categorical relationship in which the highest rank may need to be excluded to produce meaningful result. Therefore, a new binned variable named *'rank_binned'* is created as shown in the **(Figure 74)**. All in all, the final variables that will be inputted into the ordered logit model as the independent variables are: *'reviewyear_binned'*, *'ras_entrymode'*, *'award'*, *'rank_binned'*, *'rastype'*, *'supvwhite' and 'white'* as shown in the **(Figure 75)**. In the Parameter Estimates table as shown in the **(Figure 76)**, we see the coefficients, their standard errors, the Wald test, and associated p-values (Sig.), and the 95% confidence interval of the coefficients. Review year (II and IV), origin of RAS employee, and rank (I and II) are statistically significant; So for review year, we would say that if a staff belongs to the review year 2014 or 2016, we expect a -0.880 and -0.798 decrease in the ordered log odds of being in a higher review rating, given all of the other variables in the model are held constant. For origin of RAS employee, we would say that if a staff works in a pre/post-award stage, we would expect a -0.362 decrease in pre-award and -0.608 in a post-award stage in the log odds of being in a higher review rating, given that all of the other variables in the model are held constant. For rank, we would say that if a staff has a rank of I or II, we would expect a -2.217 decrease in rank I and -1.534 in rank II in the log odds of being in a higher review rating, given that all of the other variables in the model are held constant.

In the following question, we are asked to explore different factors that can explain the number of merit awards received by an employee and their rate of promotion. Variables in the *assign2_promotion_merit.sav* are similar to the previous dataset since there are identical columns such as *'uid'*, *'rank'*, *'title'*, *'award'*, *'division'*, *'department'*, *'age'*, *'gender'*, *'supvgender'*, *'rastype'*, *'ras_entrymode'*, *'white'*, *'supvwhite'*, *'educlv'*, and *'ras_treat'*. The new variables in this dataset that in not present in the previous one are *'c_merit'*, *'f_prom'*, *'t_prom'* and *'tenure_te'*. *'tenure_te'* is a numeric variable that stands for Tenure at Emory. *'ras_treat'* is a numeric variable that stands for treatment group (basic sciences & DOM). *'f_prom'* is a variable that indicates promotion (if the value is 1 and 0 otherwise). And finally *'t_prom'* is a double variable that indicates analysis time to first promotion or when record ends. To determine different factors that can explain the number of merit awards received by an employee, either Poisson regression or negative binomial regression can be considered. As you can see from the **(Figure 77)**, it is evident that the mean value (1.381) of count of merit (dependent variable) is similar to that of variance (1.908). Therefore, it is more reasonable to consider Poisson regression, which can more accurately predict a dependent variable that consists of "count data" given one or more independent variables. When analyzing data using Poisson regression, part of the process involves making sure that the data is in the right format by satisfying the four assumptions:

- Assumption #1: Your dependent variable consists of count data. count variables require integer data that must be zero or greater. Since count data must be positive value, it cannot consist of any negative value.
- Assumption #2: we can have one or more independent variables, which can be measured on a continuous, ordinal, or nominal/dichotomous scale.
- Assumption #3: You should have independence of observations. This means that each observation is independent of the other observations.
- Assumption #4: The distribution of counts (conditional on the model) follows a Poisson distribution, meaning that the observed and expected counts should be equal or at least very similar and consequently leading to the mean and variance to be almost identical.

It is always a good practice to first start by observing each variable in the dataset. When computing the descriptive statistics for each variable in this dataset, we can see the similar result as the previous result had displayed. As shown in the **(Figure 78)**, a variable *'rank'* is quite skewed; therefore, it wouldn't be a bad idea to recode this variable into a different categorical variable with more fine-grained distribution for each class. As shown in the figure **(Figure 79)**, a variable *'division'* is a categorical textual; and given the nature of the data, it cannot be entered into a model. A variable *'age'* is a continuous variable that ranges from 22.29 to 69.72; this variable can also be recoded into a categorical value to be put into a model as shown in the **(Figure 80)**. A variable *'department'* and *'rastype'* contains basically the identical set of information. Since the data within *'rastype'* is more comprehensive, the values can be recoded as the other variables. A variable *'tenure_te'* is a categorical variable that ranges from 0 to 374.24; this variable can also be recoded into a categorical value to be put into a model as shown in the **(Figure 81)**. A variable *'c_merit'* is a count variable that ranges from 0 to 6 as shown in the **(Figure 82)**. A variable *'t_prom'* is a continuous variable that contains information regarding analysis time. It is basically a time variable that can be entered as a time variable when performing survival analysis. For the sake of determining which variables are considered significant, we can perform another chi-square test after recoding these variables into the respective categorical variables.

Count of merit * Division Based on the chi-square test as shown in the **(Figure 83)**, the value of the test statistic is 70.527. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and division.

Count of merit * Department Based on the chi-square test as shown in the **(Figure 84)**, the value of the test statistic is 124.955. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and division.

Count of merit * Supervisor gender Based on the chi-square test as shown in the **(Figure 85)**, the value of the test statistic is 44.506. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and division.

Count of merit * RAS unit Based on the chi-square test as shown in the **(Figure 86)**, the value of the test statistic is 124.955. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and RAS Unit.

Count of merit * Origin of RAS employee Based on the chi-square test as shown in the **(Figure 87)**, the value of the test statistic is 28.823. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and Origin of RAS employee.

Count of merit * Treatment group Based on the chi-square test as shown in the **(Figure 88)**, the value of the test statistic is 13.344. The corresponding p-value of the test statistic is $p = 0.038$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and treatment group.

Count of merit * 1 if failure; 0 if censored. Based on the chi-square test as shown in the **(Figure 89)**, the value of the test statistic is 41.707. The corresponding p-value of the test statistic is $p < 0.001$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and f_prom.

Count of merit * Age Binned Based on the chi-square test as shown in the **(Figure 90)**, the value of the test statistic is 41.331. The corresponding p-value of the test statistic is $p = 0.015$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and age.

Count of merit * Tenure at Emory Binned Based on the chi-square test as shown in the **(Figure 91)**, the value of the test statistic is 47.789. The corresponding p-value of the test statistic is $p = 0.090$. Since the p-value is smaller than our chosen significance level ($\alpha = 0.05$), we reject the null hypothesis. Thus, we conclude that there is enough evidence to suggest an association between count of merit and tenure. Based on the chi-square test as shown in the **(Figure 92)**, only variables that are considered statistically significant will be entered into Poisson regression.

The Goodness of Fit table shown in the **(Figure 93)** provides many measures that can be used to assess how well the model fits. When looking at the value in the *"Value/df"* column for the *"Pearson Chi-Square"* row, which is 0.803 in this case, as shown above: A value of 1 indicates equi-dispersion whereas values greater than 1 indicate overdispersion and values below 1 indicate under-dispersion. The Omnibus Test table shown in the **(Figure 94)** shows a likelihood ratio test of whether all the independent variables collectively improve the model over the intercept-only model (i.e., with no independent variables added). Having all the independent variables in our example model we have a p-value that is less than .001 (i.e., $p < .001$), indicating a statistically very significant overall model, as shown below in the "Sig." column. The Tests of Model Effects table as shown in the **(Figure 95)** displays the statistical significance of each of the independent variables in the "Sig." column: There is not usually any interest in the model intercept. However, we can see that the supervisor gender ($p = .450$) and tenure at Emory ($p = .505$) were not statistically significant, but the origin of RAS employee ($p = 0.015$), 1 if failure; 0 if censored ($p < 0.001$), and age ($p = 0.045$) are statistically significant. This table is mostly useful for categorical independent variables because it is the only table output that considers the overall effect of a categorical variable. This table shown in the **(Figure 96)** provides both the coefficient estimates (the "B" column) of the Poisson regression and the exponentiated values of the coefficients (the "Exp(B)" column). It is usually the latter that are more informative. Consider, for example, the origin of RAS employee (where the value equals 1). The exponentiated value is 0.500. This means that the number of merit awards (i.e., the count of the dependent variable) will be 0.500 times bigger for those who are from different RAS division. Another way of saying this is that there is a 50% decrease in the number of publications for those who are from different RAS division. Consider, another example, the department (SOM = Neuroscience). The exponentiated value is 0.349. This means that the number of merit awards (i.e., the count of the dependent variable) will be 0.349 times bigger for those who are from neuroscience department. Another way of saying this is that there is about 65% decrease in the number of merit awards for those who are from neuroscience department. A similar interpretation can be made for the categorical variable.

Next, we are now exploring different factors that can explain the rate of promotion for employees at Emory. We need to perform survival analysis in forms of Kaplan-Meier Estimates and Cox regression. We should first conduct log-rank test to determine the variables that are statistically valid as shown in the **(Figure 97)**. Based on the **(Figure 98)**, we can conduct log-rank test by going to the Kaplan-Meier section under Survival Analysis. As you can see from the table shown in the **(Figure 99)**, it is evident that none of the variables is statistically significant given the p-value that is greater than 0.05. The only variable that is considered significant is the count of merit, and Kaplan-Meier plot is also shown in the **(Figure 100)**. Next, we can compute life table. Looking at the descriptive statistics of *'t_prom'* variable as shown in the **(Figure 101)**, we can see that the value ranges from 0.26 to 66.15. And we can our number accordingly when enter the time intervals. The table shown on the **(Figure 102)** shows the life table. The **(Figure 103)** displays the survival plot. Finally, it is the time to perform cox regression. Time, status, and covariates variables can be entered as shown in the **(Figure 104)**. It is important to note that categorical covariates are also entered to output a proper result. As seen in the **(Figure 105)**, Regression coefficient (B) – positive beta coefficient means that those who receive more merit awards are also likely to experience higher rates of promotion. Hazard ratio (exp(B)) provides a more interpretable effect size of the covariates. Receiving merit awards increases the hazard (rates of promotion) by factor of 2.304 or 130.4 percent. Other variables need not interpretation since p-value is greater than 0.05.