University of Illinois at Urbana-Champaign

Salary Analysis

Team 32

Jinna Yoon (Contributions: Introduction, Association Tests, Logistic Regression, Model Evaluation, Conclusion)

Carol Gong (Contributions: Data Cleaning, Exploratory Data Analysis, Feature Selection, Discussion)

Advanced Data Analysis (STAT 448)

Lelys Bravo de Guenni

12/15/24

Introduction

Salary is an important factor in determining the desirability of a job. A lot of students choose their career path and future investments into obtaining a well-paying job so that they can live comfortably. However, there are many factors that go into determining what jobs they want and how to pursue it.

Data description: This dataset is the "Census Income" dataset from the UC Irvine Machine Learning Repository. The data was extracted by Barry Becker from the 1994 Census database. In this dataset, we can see that the comma-separated file has categories for Age, WC (Work Class), Education, Ednum (Education Number or number of years spent for an education), MS (marital-status), Occupation, Relationship, Race, Sex, Hpw (hours-per week worked), and NCountry (native-country that person is working from). Each row is a person.

Goals: In this project, we will create a model that will determine which factors predict salary the best so that students can see which factors can help increase salary. Students can prepare the steps they need to take in the future to obtain that level of salaried job. We would also like to perform model evaluation to ensure that the resulting model can optimally predict salary. In other words, more explicitly:

1. Which of the variables in the salary dataset can explain the level of salary best?
2. Is our model (logistic regression) the best model for this data?

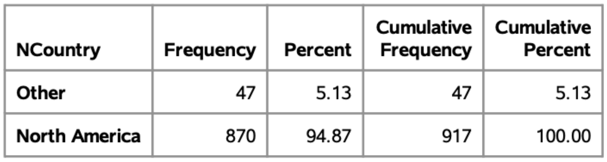
Data Cleaning

Before performing data analysis, we cleaned up the data by merging levels within most of the variables to create larger groups. For each variable, we made sure to have 5 levels maximum, as many of them had very small numbers of observations. We wanted to make sure each level had at least 20 observations to conduct analysis. In some of the groupings, we added additional context to the levels that were not present in the initial dataset. For example, we created subcategories for Age with specified ranges to define what constitutes as "elderly" and "young adults". We also did this with Hpw (working hours per week) by separating this variable into part-time and full-time schedules. It is important to note that in doing so, we converted Age, Ednum, and Hpw from a continuous variable to a categorical one by making subcategories. For the other categorical variables in the dataset, we simply grouped similar levels together. For example, we reduced the number of original levels that were in NCountry by grouping countries by what continent they were in. Due to the small observations in countries not residing in North America, NCountry resulted in having only two levels – North America and "Other". "Other" included countries that combined their few observations into a greater number of observations that can be used in future analysis.

Exploratory Data Analysis

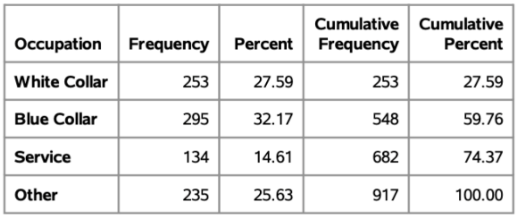
To conduct exploratory data analysis, we produced frequency tables[1] and ran summary statistics for all variables. Looking at the frequency tables, there is a trend across many variables where some levels have many more observations than other levels. This inequality is most apparent with NCountry (Figure 1.1), where North America has many more observations than all other continents combined. There are a few variables in which the number of observations are more equally distributed across all levels: for example, MS (Figure 1.2) and Occupation (Figure 1.3) are more equal in all levels than the other variables.

Figure 1.1



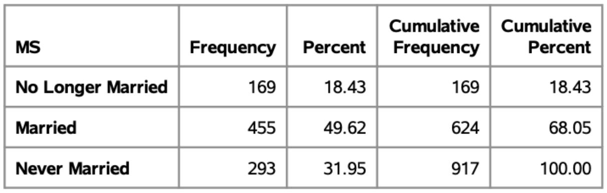
NCountry	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Other	47	5.13	47	5.13
North America	870	94.87	917	100.00

Figure 1.2



Occupation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
White Collar	253	27.59	253	27.59
Blue Collar	295	32.17	548	59.76
Service	134	14.61	682	74.37
Other	235	25.63	917	100.00

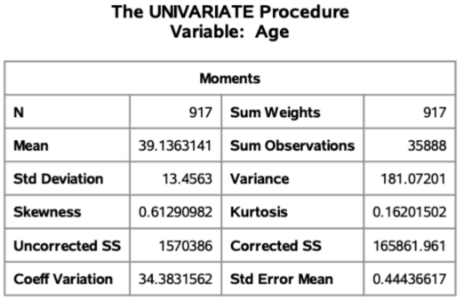
Figure 1.3



MS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No Longer Married	169	18.43	169	18.43
Married	455	49.62	624	68.05
Never Married	293	31.95	917	100.00

Next, to look at the summary statistics for quantitative variables, we will go over the mean, standard deviation, and skewness of the variables. First, for Age (Figure 1.4), the mean is 39.136 years (which is classified as early middle age), the standard deviation is 13.456 years, and the skewness is 0.613 (this indicates a moderate skew to the right). For Ednum (Figure 1.5), the mean years of education is 10.092, the standard deviation is 2.569, and the skewness is -0.383 (which is a moderately weak left skew). Finally, for Hpw (Figure 1.6), or hours worked weekly, the mean is 40.531, the standard deviation is 11.647, and the skewness is 0.010 (which is a very weak right skew).

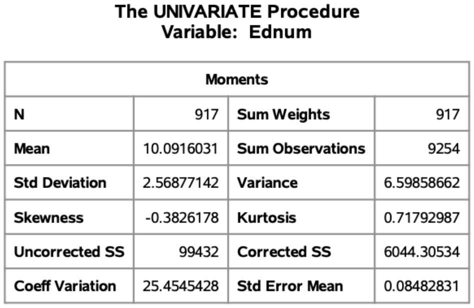
Figure 1.4



The UNIVARIATE Procedure
Variable: Age

Moments			
N	917	Sum Weights	917
Mean	39.1363141	Sum Observations	35888
Std Deviation	13.4563	Variance	181.07201
Skewness	0.61290982	Kurtosis	0.16201502
Uncorrected SS	1570386	Corrected SS	165861.961
Coeff Variation	34.3831562	Std Error Mean	0.44436617

Figure 1.5

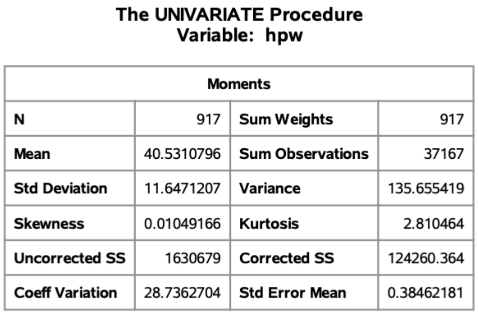


The UNIVARIATE Procedure
Variable: Ednum

Moments			
N	917	Sum Weights	917
Mean	10.0916031	Sum Observations	9254
Std Deviation	2.56877142	Variance	6.59858662
Skewness	-0.3826178	Kurtosis	0.71792987
Uncorrected SS	99432	Corrected SS	6044.30534
Coeff Variation	25.4545428	Std Error Mean	0.08482831

[1] Frequency tables for exploratory data analysis will appear in Section ii of Appendix A

Figure 1.6



The UNIVARIATE Procedure
Variable: hpw

Moments			
N	917	Sum Weights	917
Mean	40.5310796	Sum Observations	37167
Std Deviation	11.6471207	Variance	135.655419
Skewness	0.01049166	Kurtosis	2.810464
Uncorrected SS	1630679	Corrected SS	124260.364
Coeff Variation	28.7362704	Std Error Mean	0.38462181

Along with summary statistics and frequency tables, we also generated histograms and probability plots for the quantitative variables. For this part only, we removed the grouping of the variables' levels so that the graphs can be displayed and analyzed optimally (however, we can describe the histograms in particular in relation to the levels in the text). First, looking at the histogram for Age (Figure 1.7), we can see that there is a right skew to the graph, and it peaks at around 30, which is the young adults category. Then, looking at the probability plot (Figure 1.8), we can see that it generally follows a diagonal line, except at the ends where it curls. This confirms that the data is not normally distributed. Then, with Ednum, we can see that in the histogram (Figure 1.9), 9 years of education is the most common by far, followed by 10 and 13, and there is a skew towards the left. The probability plot (Figure 1.10) follows a stepwise pattern that deviates from a diagonal line, also suggesting that the variable is not normally distributed. Finally, with Hpw, the number of hours worked per week peaks at around 39 according to the histogram (Figure 1.11), which can be categorized as a full-time work schedule. The graph also skews right (thought very slight). The probability plot (Figure 1.12) follows the same stepwise pattern as Ednum, but it veers off the diagonal line more dramatically, indicating that it is not normally distributed.

Figure 1.7

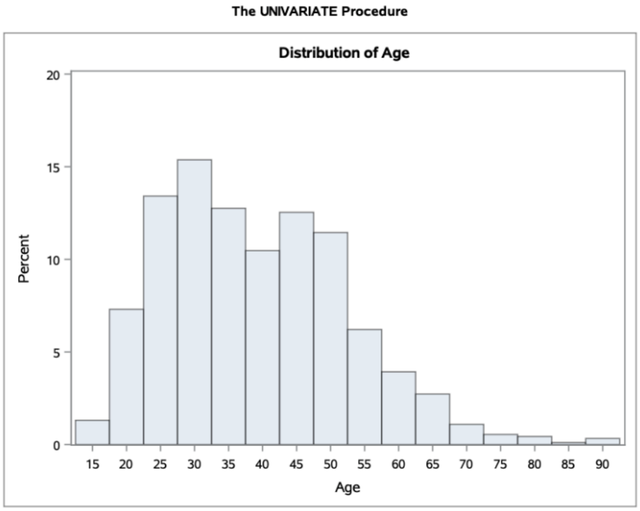


Figure 1.8

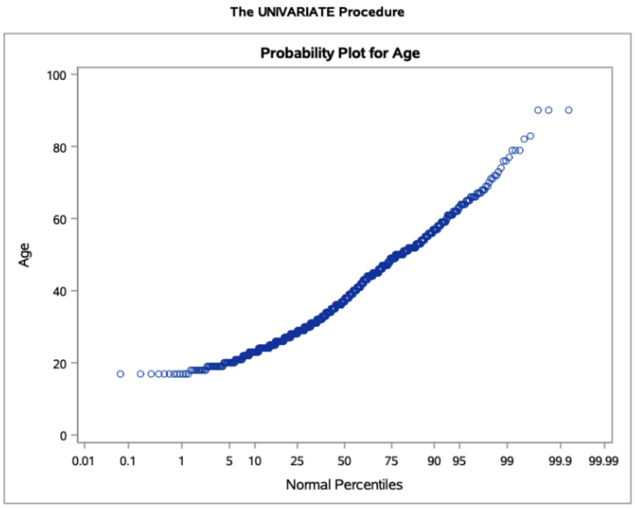


Figure 1.9

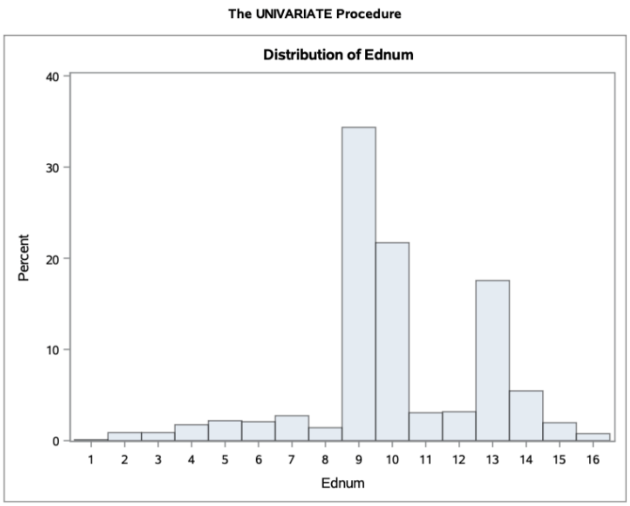


Figure 1.10

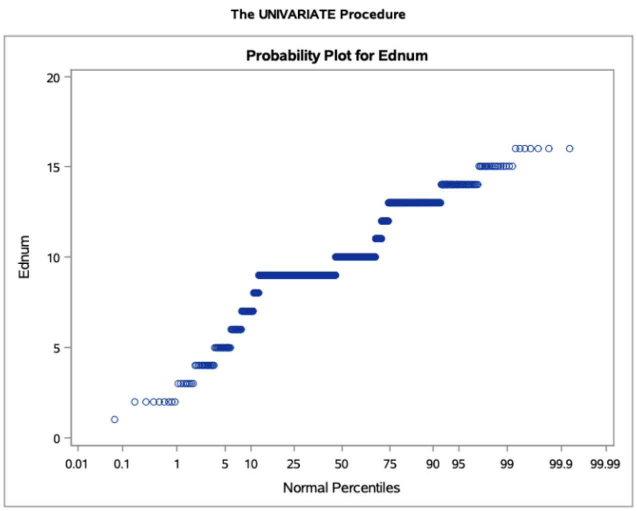


Figure 1.11

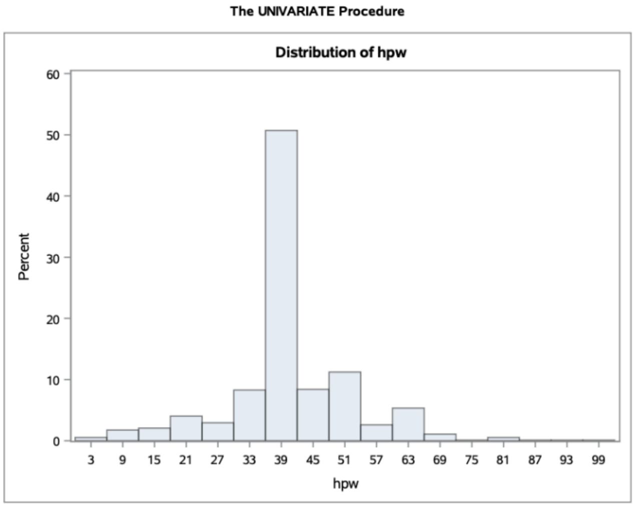
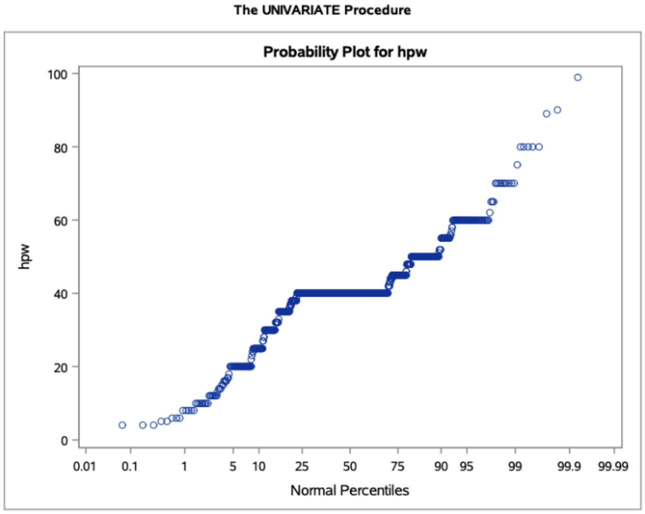


Figure 1.12

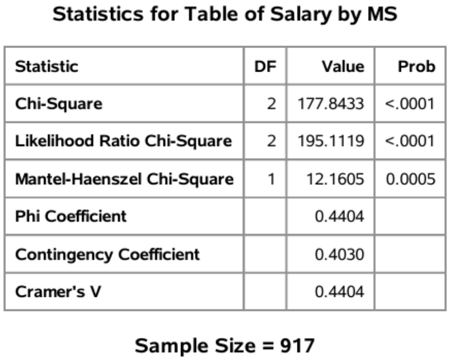


Modeling

i. Association Tests

We decided to look at each of the variables' association with Salary using association tests. For all tests, we will be using a significance threshold of 0.05. The contingency table[2] for Salary and MS have most of the expected frequencies in each cell > 5. This allows us to use the chi-square association test (Figure 2.1), where we have p-values that are < 0.0001. Because this p-value is less than our significance threshold of 0.05, we reject the test's null hypothesis and conclude that there is evidence for an association between Salary and MS.

Figure 2.1

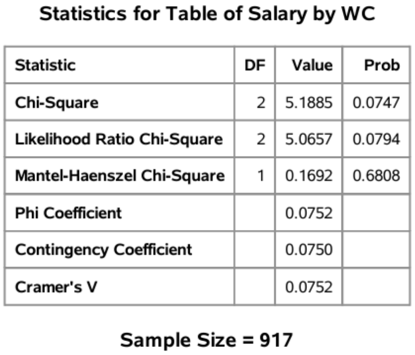


Statistics for Table of Salary by MS

Statistic	DF	Value	Prob
Chi-Square	2	177.8433	<.0001
Likelihood Ratio Chi-Square	2	195.1119	<.0001
Mantel-Haenszel Chi-Square	1	12.1605	0.0005
Phi Coefficient		0.4404	
Contingency Coefficient		0.4030	
Cramer's V		0.4404	

Sample Size = 917

Figure 2.2



Statistics for Table of Salary by WC

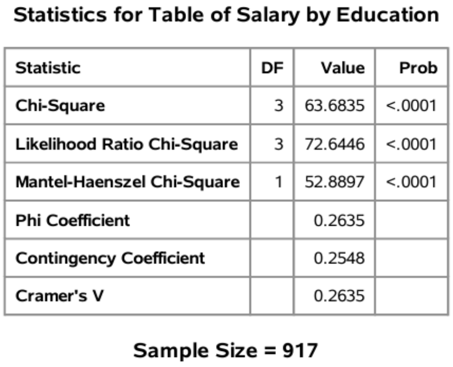
Statistic	DF	Value	Prob
Chi-Square	2	5.1885	0.0747
Likelihood Ratio Chi-Square	2	5.0657	0.0794
Mantel-Haenszel Chi-Square	1	0.1692	0.6808
Phi Coefficient		0.0752	
Contingency Coefficient		0.0750	
Cramer's V		0.0752	

Sample Size = 917

The contingency table for Salary and WC has all of the expected frequencies in each cell > 5. We use the chi-square association test (Figure 2.2), which has a p-value of 0.0747. We fail to reject our null hypothesis and conclude that there is evidence for no association between Salary and WC.

The contingency table for Salary and Education have majority of the expected frequencies in each cell > 5. We use the chi-square association test (Figure 2.3), which has a p-value < 0.0001. We reject the null hypothesis and conclude that there is evidence for an association between Salary and Education.

Figure 2.3

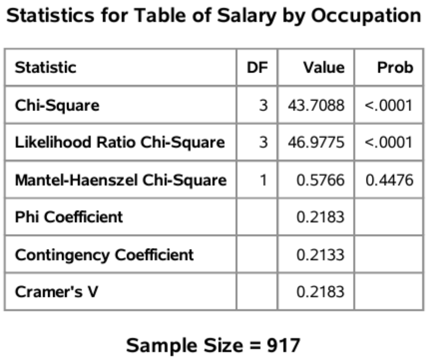


Statistics for Table of Salary by Education

Statistic	DF	Value	Prob
Chi-Square	3	63.6835	<.0001
Likelihood Ratio Chi-Square	3	72.6446	<.0001
Mantel-Haenszel Chi-Square	1	52.8897	<.0001
Phi Coefficient		0.2635	
Contingency Coefficient		0.2548	
Cramer's V		0.2635	

Sample Size = 917

Figure 2.4



Statistics for Table of Salary by Occupation

Statistic	DF	Value	Prob
Chi-Square	3	43.7088	<.0001
Likelihood Ratio Chi-Square	3	46.9775	<.0001
Mantel-Haenszel Chi-Square	1	0.5766	0.4476
Phi Coefficient		0.2183	
Contingency Coefficient		0.2133	
Cramer's V		0.2183	

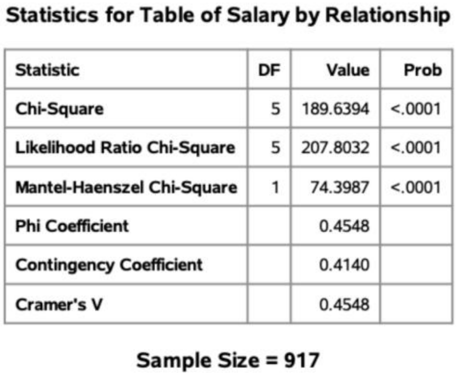
Sample Size = 917

[2] All contingency tables will be listed in Section i in Appendix A

The contingency table for Salary and Occupation has expected frequencies for all the cells > 5. We use the chi-square association test (Figure 2.4), which has a p-value < 0.0001. We reject the null hypothesis and conclude that there is evidence for an association between Salary and Occupation.

The contingency table for Salary and Relationship has expected frequencies for all cells > 5. We use the chi-square association test (Figure 2.5), which has a p-value < 0.0001. We reject the null hypothesis and conclude that there is an association between Salary and Relationship.

Figure 2.5

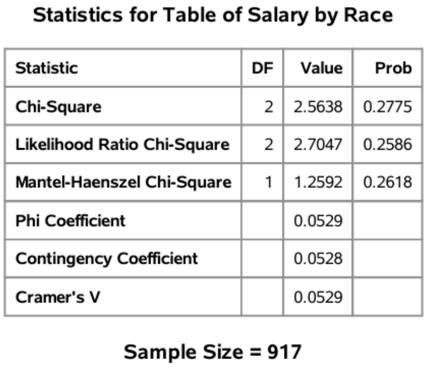


Statistics for Table of Salary by Relationship

Statistic	DF	Value	Prob
Chi-Square	5	189.6394	<.0001
Likelihood Ratio Chi-Square	5	207.8032	<.0001
Mantel-Haenszel Chi-Square	1	74.3987	<.0001
Phi Coefficient		0.4548	
Contingency Coefficient		0.4140	
Cramer's V		0.4548	

Sample Size = 917

Figure 2.6



Statistics for Table of Salary by Race

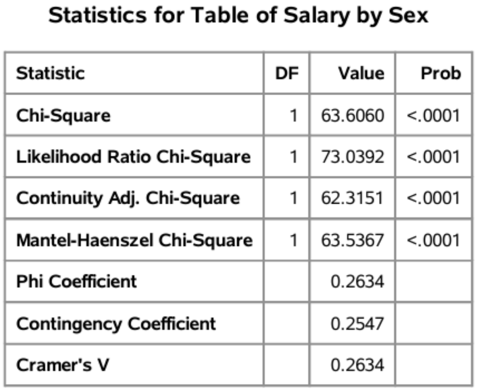
Statistic	DF	Value	Prob
Chi-Square	2	2.5638	0.2775
Likelihood Ratio Chi-Square	2	2.7047	0.2586
Mantel-Haenszel Chi-Square	1	1.2592	0.2618
Phi Coefficient		0.0529	
Contingency Coefficient		0.0528	
Cramer's V		0.0529	

Sample Size = 917

The contingency table for Salary and Race has expected frequencies for all cells > 5. We use the chi-square association test (Figure 2.6), which has a p-value of 0.2775. We fail to reject the null hypothesis and conclude that there is no association between Salary and Race.

The contingency table for Salary and Sex has expected frequencies in all of the cells > 5. We use the chi-square association test (Figure 2.7), which gives us a p-value < 0.0001. We reject the null hypothesis and conclude that there is evidence for an association between Salary and Sex.

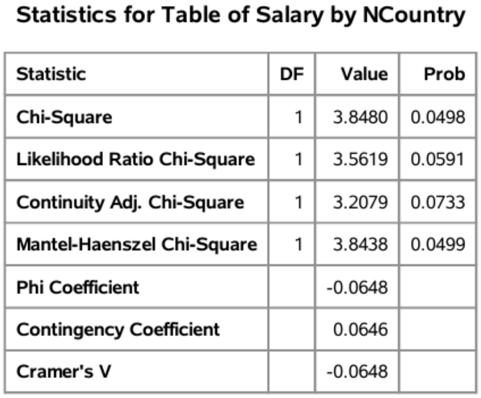
Figure 2.7



Statistics for Table of Salary by Sex

Statistic	DF	Value	Prob
Chi-Square	1	63.6060	<.0001
Likelihood Ratio Chi-Square	1	73.0392	<.0001
Continuity Adj. Chi-Square	1	62.3151	<.0001
Mantel-Haenszel Chi-Square	1	63.5367	<.0001
Phi Coefficient		0.2634	
Contingency Coefficient		0.2547	
Cramer's V		0.2634	

Figure 2.8



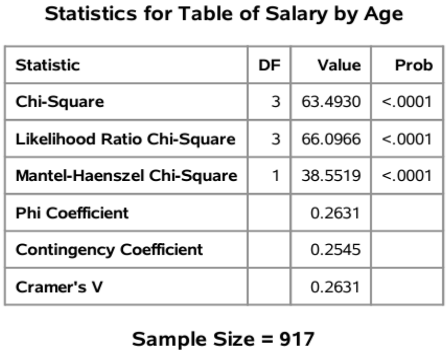
Statistics for Table of Salary by NCountry

Statistic	DF	Value	Prob
Chi-Square	1	3.8480	0.0498
Likelihood Ratio Chi-Square	1	3.5619	0.0591
Continuity Adj. Chi-Square	1	3.2079	0.0733
Mantel-Haenszel Chi-Square	1	3.8438	0.0499
Phi Coefficient		-0.0648	
Contingency Coefficient		0.0646	
Cramer's V		-0.0648	

The contingency table for Salary and NCountry has majority of the expected frequencies in all cells > 5. We use the chi-square association test (Figure 2.8), which gives us a p-value of 0.0498. We reject the null hypothesis and conclude that there is evidence for an association between Salary and NCountry.

The contingency table for Salary and Age has majority of the expected frequencies in all cells > 5. We use the chi-square association test (Figure 2.9), which gives us a p-value < 0.0001. We reject the null hypothesis and conclude that there is evidence for an association between Salary and Age.

Figure 2.9

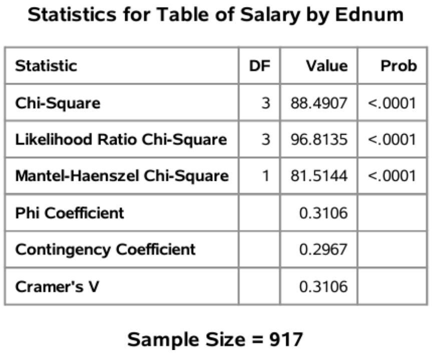


Statistics for Table of Salary by Age

Statistic	DF	Value	Prob
Chi-Square	3	63.4930	<.0001
Likelihood Ratio Chi-Square	3	66.0966	<.0001
Mantel-Haenszel Chi-Square	1	38.5519	<.0001
Phi Coefficient		0.2631	
Contingency Coefficient		0.2545	
Cramer's V		0.2631	

Sample Size = 917

Figure 2.10



Statistics for Table of Salary by Ednum

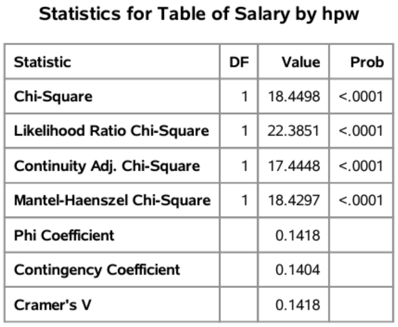
Statistic	DF	Value	Prob
Chi-Square	3	88.4907	<.0001
Likelihood Ratio Chi-Square	3	96.8135	<.0001
Mantel-Haenszel Chi-Square	1	81.5144	<.0001
Phi Coefficient		0.3106	
Contingency Coefficient		0.2967	
Cramer's V		0.3106	

Sample Size = 917

The contingency table for Salary and Ednum has majority of the expected frequencies in all cells > 5. We use the chi-square association test (Figure 2.10), which gives us a p-value < 0.0001. We reject the null hypothesis and conclude that there is evidence for an association between Salary and Ednum.

The contingency table below for Salary and Hpw has majority of the expected frequencies in all cells > 5. We use the chi-square association test (Figure 2.11), which gives us a p-value < 0.0001. We reject the null hypothesis and conclude that there is evidence for an association between Salary and Hpw.

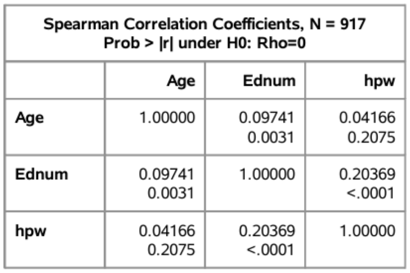
Figure 2.11



Statistics for Table of Salary by hpw

Statistic	DF	Value	Prob
Chi-Square	1	18.4498	<.0001
Likelihood Ratio Chi-Square	1	22.3851	<.0001
Continuity Adj. Chi-Square	1	17.4448	<.0001
Mantel-Haenszel Chi-Square	1	18.4297	<.0001
Phi Coefficient		0.1418	
Contingency Coefficient		0.1404	
Cramer's V		0.1418	

Figure 2.12



Spearman Correlation Coefficients, N = 917
Prob > |r| under H0: Rho=0

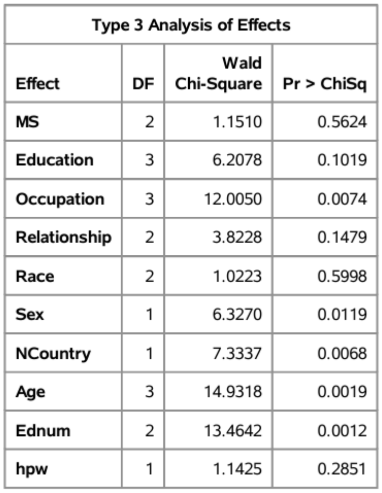
	Age	Ednum	hpw
Age	1.00000	0.09741 0.0031	0.04166 0.2075
Ednum	0.09741 0.0031	1.00000	0.20369 <.0001
hpw	0.04166 0.2075	0.20369 <.0001	1.00000

To ensure that there is no multicollinearity between the quantitative variables, a correlation matrix was created. In Figure 2.12, a Spearman correlation table was created since the normality assumption was not fulfilled for all three variables (refer back to the exploratory data analysis). We can see that all combinations of Age, Ednum, and Hpw have very small correlation values that are less than 0.21, indicating that there is little to weak correlation amongst these variables. There are no multicollinearity issues.

ii. Logistic Regression

We can identify that salary is a binary categorical variable that is sorted either <= 50k or >50k. To explain salary, a two level categorical, as a relationship with the other variables in the salary dataset, we would have to fit a logistic regression model. After finding which variables may have an association with Salary, we would like to fit them as a function for Salary. The Type 3 Analysis of Effects table (Figure 3.1) shows that Occupation, Sex, NCountry, Age, and Ednum are statistically significant predictors.

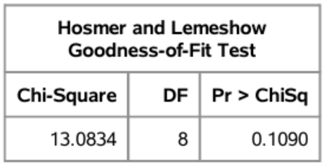
Figure 3.1



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
MS	2	1.1510	0.5624
Education	3	6.2078	0.1019
Occupation	3	12.0050	0.0074
Relationship	2	3.8228	0.1479
Race	2	1.0223	0.5998
Sex	1	6.3270	0.0119
NCountry	1	7.3337	0.0068
Age	3	14.9318	0.0019
Ednum	2	13.4642	0.0012
hpw	1	1.1425	0.2851

Figure 3.2

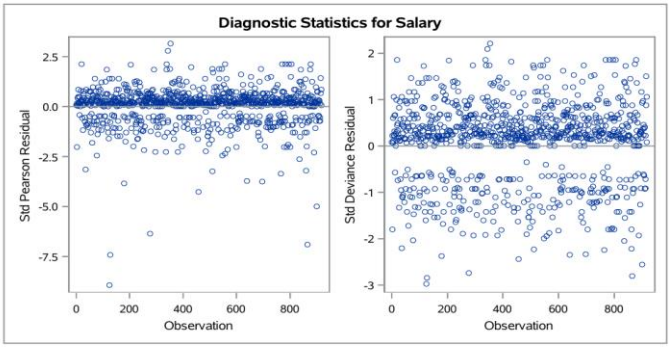


Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
13.0834	8	0.1090

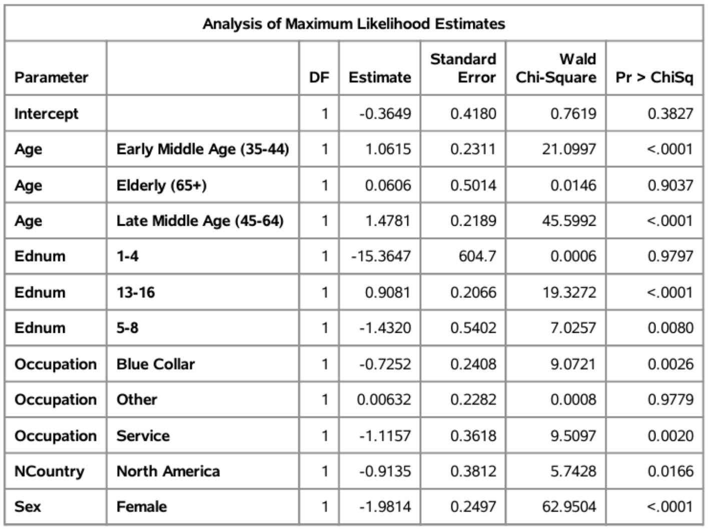
The Hosmer and Lemeshow Goodness of Fit Test (Figure 3.2) has a p-value of 0.109. This means that we would fail to reject the null hypothesis that assumes the model is a good fit, providing evidence that this model is not lack of fit. However, this could occur due to having too many predictors in our model. To see if there are any diagnostic problems with this logistic regression model, Pearson and Deviance residual plots (Figure 3.3) were created. The Pearson residual plot shows that this model may have a handful of outliers or values that were not able to be accurately predicted by this model, but the majority fall on the 0 horizontal line which shows that the model can predict the majority of the data points fairly well. The deviance residual plot has values that mostly range from -2 to 2. This range is not that big, providing evidence that our model does not have overdispersion.

Figure 3.3



iii. Feature Selection

To find the best logistic regression model, feature selection was performed to reduce the number of current variables and select the predictors that can best explain Salary. After conducting stepwise selection, we end up having Occupation, Sex, NCountry, Age, and Ednum as significant predictors that help explain Salary. On a level-specific interpretation, Age [Early Middle Age (35-44)], Age [Late Middle Age (45-64)], Ednum [13-16], Ednum [5-8], Occupation [Blue Collar], Occupation [Service], NCountry[North America], and Sex [Female].



Analysis of Maximum Likelihood Estimates

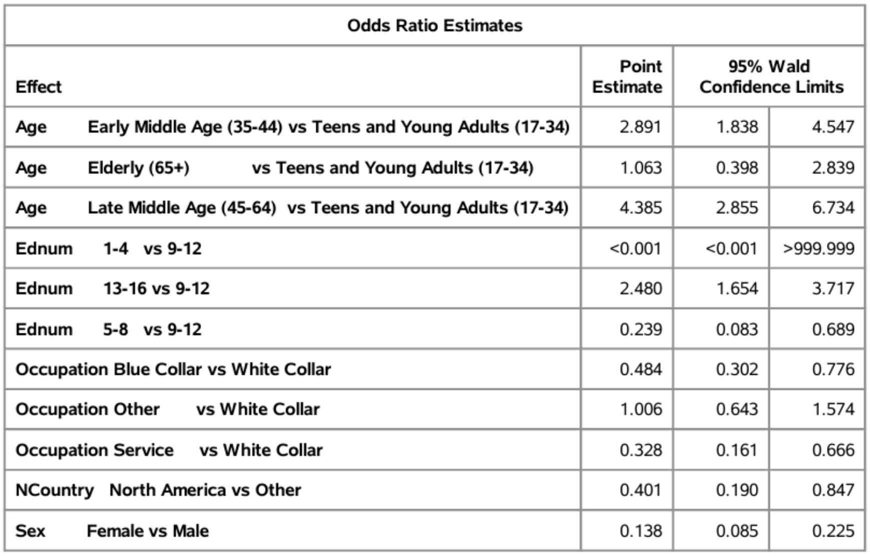
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.3649	0.4180	0.7619	0.3827
Age	Early Middle Age (35-44)	1	1.0615	0.2311	21.0997	<.0001
Age	Elderly (65+)	1	0.0606	0.5014	0.0146	0.9037
Age	Late Middle Age (45-64)	1	1.4781	0.2189	45.5992	<.0001
Ednum	1-4	1	-15.3647	604.7	0.0006	0.9797
Ednum	13-16	1	0.9081	0.2066	19.3272	<.0001
Ednum	5-8	1	-1.4320	0.5402	7.0257	0.0080
Occupation	Blue Collar	1	-0.7252	0.2408	9.0721	0.0026
Occupation	Other	1	0.00632	0.2282	0.0008	0.9779
Occupation	Service	1	-1.1157	0.3618	9.5097	0.0020
NCountry	North America	1	-0.9135	0.3812	5.7428	0.0166
Sex	Female	1	-1.9814	0.2497	62.9504	<.0001

The equation of the model with the significant predictors would be interpreted as:

Log odds of Salary = -0.3649 + 1.0615 * Early Middle Age + 1.4781 * Late Middle Age - 1.4320 * Ednum (5-8) + 0.9081 * Ednum (13-16) - 0.7252 * Blue Collar + 0.00632 * Other - 1.1157 * Service - 0.9135 * North America - 1.9814 * Female

This model provides information on the log odds of someone making a >50k salary based on occupation, sex, country of origin, age, and number of years of education in the figure below. For each variable, the log odds are higher or lower at a certain level, holding other variables constant and relative to the baseline level. For the point estimates for the log odds, a value over 1 increases the log odds, while a value under 1 decreases the log odds. Only the interpretation of significant odds ratios will be included, which can be determined if 1 is not in the confidence interval. Looking at the table, there are 8 odds ratios that are significant, which are Age Early Middle Age (35-44) vs Teens and Young Adults (17-34), Age Late Middle Age (45-64) vs Teens and Young Adults (17-34), Ednum 13-16 vs 9-12, Ednum 5-8 vs 9-12, Occupation Blue Collar vs White Collar, Occupation Service vs White Collar, NCountry North America vs Other, and Sex Female vs Male. For early middle age vs. teens and young adults, early middle age individuals are 2.891 times more likely to earn a >50k salary compared to young adults and teens. The individual odds ratios will be interpreted below.

For late middle age vs. teens and young adults, middle age individuals are 4.385 times more likely to earn >50k. For Ednum 13-16 vs 9-12, those with 13-16 years of education are 2.480 times more likely to earn>50k compared to those with 9-12 years of education. For Ednum 5-8 vs 9-12, those with 5-8 years of education are 0.239 times less likely to earn >50k compared to those with 9-12 years of education. For blue collar vs white collar, blue collar workers are 0.484 times less likely to earn >50k compared to white collar workers. For occupation service vs. white collar, service workers are 0.328 times less likely to earn >50k compared to white collar workers. For North America vs. other, those with a native country in North America are 0.401 times less likely to make >50k compared to those whose native country are not. For female vs male, women are 0.138 times less likely to make >50k compared to men.



Odds Ratio Estimates

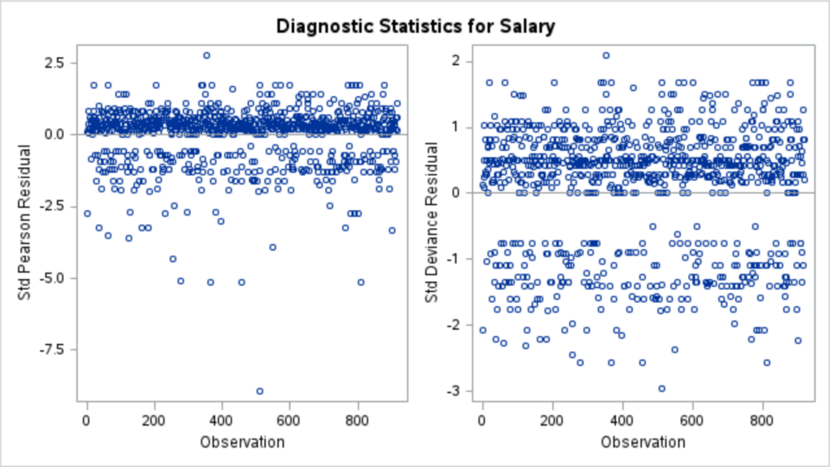
Effect	Point Estimate	95% Wald Confidence Limits	
Age Early Middle Age (35-44) vs Teens and Young Adults (17-34)	2.891	1.838	4.547
Age Elderly (65+) vs Teens and Young Adults (17-34)	1.063	0.398	2.839
Age Late Middle Age (45-64) vs Teens and Young Adults (17-34)	4.385	2.855	6.734
Ednum 1-4 vs 9-12	<0.001	<0.001	>999.999
Ednum 13-16 vs 9-12	2.480	1.654	3.717
Ednum 5-8 vs 9-12	0.239	0.083	0.689
Occupation Blue Collar vs White Collar	0.484	0.302	0.776
Occupation Other vs White Collar	1.006	0.643	1.574
Occupation Service vs White Collar	0.328	0.161	0.666
NCountry North America vs Other	0.401	0.190	0.847
Sex Female vs Male	0.138	0.085	0.225

iv. Model Evaluation

To determine if the selected logistic model of Salary explained by Occupation, Sex, NCountry, Age, and Ednum is a good model to predict future Salary, we will be looking at the residual plots (Figure 5.1), ROC curve (Figure 5.2), and confusion matrix (Figure 5.3).

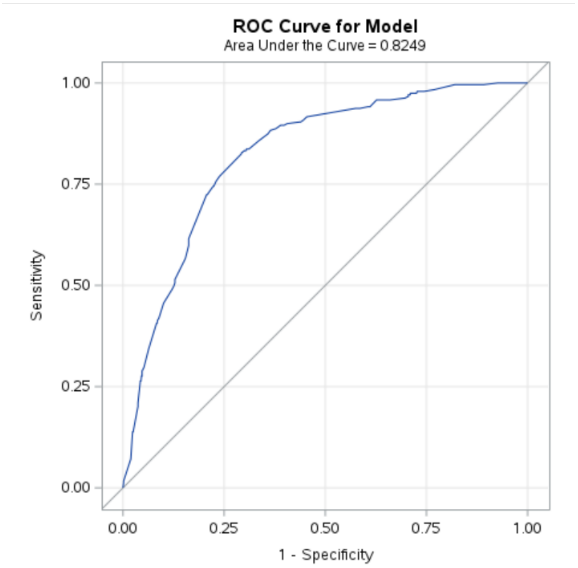
The Pearson residual plot shows that this model may have some outliers or values that were not able to accurately predicted by this model, but majority fall on the 0 horizontal line or within a 2.5 range of it. Comparing this with the full model of associated variables with Salary, this Pearson plot has a similar number of outliers and observations that were not predicted as well. The range of the residuals is also -2.5 to 2.5. This plot shows that the model can predict majority of the data points fairly well. The deviance residual plot has values that mostly range from -2 to 2, similar to the full model of associated variables. This range is not that big, providing evidence that our model does not have overdispersion.

Figure 5.1



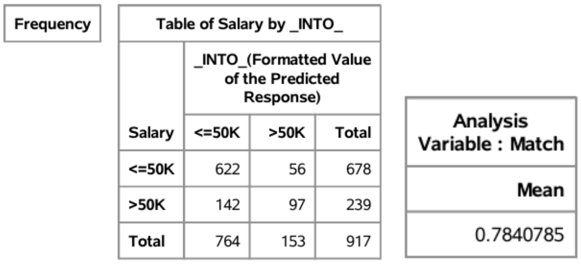
The ROC curve below indicates how well the model classifies cases of Salary <=50k and Salary >50k. The area under the curve is 0.8249. This means that our logistic model performs well at distinguishing where the data should fall between our two levels of Salary.

Figure 5.2



The table below is a confusion matrix (Figure 5.3) that displays the actual-by-predicted classification to see if the model is actually classifying the cases to their proper Salary level. The accuracy of the classification is a 78.4% match of the actual values equal to the predicted values, which is a fair amount to classify correctly. However, Salary >50k got incorrectly classified into <=50k more times than it was correctly classified as >50k. This creates some sign of concern and prompts us to find a better model.

Figure 5.3



Frequency

Table of Salary by _INTO_

	INTO(Formatted Value of the Predicted Response)		
Salary	**<=50K**	**>50K**	**Total**
<=50K	622	56	678
>50K	142	97	239
Total	764	153	917

Analysis Variable : Match
Mean
0.7840785

Discussion

- Our goal in this report was to identify which characteristics help determine a higher salary level. We also wanted to evaluate if the model that predicted the level of Salary was a good fit.
- In our association tests, we have determined which variables have a significant relationship with Salary. Based on our results, we were able to find that MS (marriage status), Education, Occupation, Relationship, Sex, NCountry (native country), Ednum, Hpw, and Age are associated with Salary. However, while these factors may be associated with Salary, we need to see which variables are statistically correlated with Salary. With salary being a binary category variable (either <=50k or >50k), we fit a logistic model with all variables in the dataset. After performing logistic regression, it resulted in a model with Occupation, Sex, NCountry, Age, and Ednum as significant predictors. This means that these variables have a statistically significant relationship with explaining Salary. This indicates that these 5 variables are the best in explaining the level of salary from this dataset. The Hosmer and Lemeshow Goodness of Fit Test also provided evidence that this model was a good fit.
- To make sure that our model was parsimonious, we conducted feature selection. After performing stepwise selection, our final model included Occupation, Sex, NCountry, Age, and Ednum as significant predictors for Salary. By looking at the residual plots, ROC curve, and confusion matrix, we were able to determine that the model predicts the level of Salary fairly well. We also verified if the model properly classified whether an observation falls into the <=50k or >50k categories. The confusion matrix (Figure 5.3) further confirms this, giving us a 78.4% match with predicted and actual salaries. However, while this model did perform generally well, there are still areas for concern. As noted in the confusion matrix, there was a disparity in misclassification of <= 50k and >50k salaries, with rates being higher for salaries <= 50k. This indicates that the model performs more poorly when predicting lower salaries.
- For our resulting significant factors, it intuitively makes sense that early middled aged or late middled aged employees make significantly more than the rest of the other age groups. Tenured employees tend to have more experience and stability in their careers that oftentimes results in higher pay. Those with higher years of education have higher odds of obtaining a higher level salary. Many specialized jobs (grouped as Other in Occupation or White Collar in Occupation) have higher pay due to the high skill requirement. These skills and knowledge that is needed is often obtained through advanced education. It would be most typical for someone to attend college or graduate school if their Ednum exceeds. Therefore, it is perfectly reasonable for Ednum [13-16] to have higher odds of obtaining a higher level of Salary compared to the lesser education number groups. Interestingly, countries in North America have less odds of having a higher level of salary compared to other countries. This could be due to poorer countries that are inside of the continent (which we grouped) such as Mexico, Cuba, El-Salvador, Guatemala, Haiti, and Jamaica. In most countries, pay for women (female) are typically less than men. Therefore, it would make sense for sex to be significant in determining salary level – and for the odds of a higher salary decreasing due to sex as female.

- Limitations
    - While analyzing our data, there were some limitations, particularly with the data itself. While we wanted to take a less generalized approach and analyze the specific levels further, some levels had such few observations that keeping the levels as they were originally would affect our analysis negatively. This is true particularly for NCountry, where the number of observations were disproportionally from North America, and we originally wanted to analyze the data by continent. Furthermore, the countries that were grouped for North America may have been considered to be grouped in another way. Due to time constraints, we were not able to look at how each individual factor interacted with each other. Also, since this data was collected in 1994, it is outdated, and will not reflect current demographic, educational, and occupational trends. So, our analysis may not be the best fit for current university students. Lastly, in terms of the original context of our research question, some of the variables we determined that were the best predictors of Salary cannot be controlled, particularly NCountry and Sex. So, students cannot prepare in those areas to acquire a high-paying job.

Conclusion

- Using this dataset, we discovered that Salary is explained as a function of Occupation, Sex, NCountry, Age, and Ednum. The odds of obtaining a higher salary is increased when the person has a white collar or other specialized occupation, in their early or late middle ages, has at least twelve years of education, does not live in North America, and is a man.

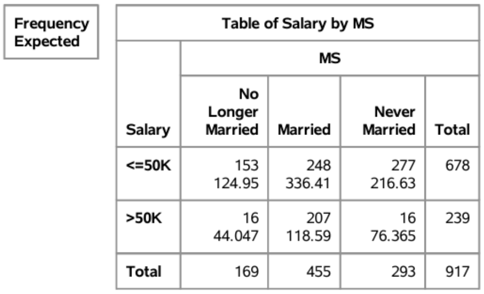
  Demographics, such as age, native country, and sex, are harder to control, but they give an idea of the characteristics of what groups expectedly earn more. To obtain the highest odds of a higher salary, they would have to obtain at least a 4-year college degree and pursue a white-collar or equivalent in specialization. If this student continues to hold onto this type of job in their career, they will expect an increase in salary as the years go on – with a peak in salary occurring when they are middle–aged.

- For future research, we would like to expand the scope of our model to include interaction terms. If possible, it would help with relevancy to obtain a model that is closer to the current year, 2024. In this way, we can generalize more toward the current student population and find relevant job trends that may have changed over the course of 30 years. For further analysis, one could also compare the model we curated with the 1994 data with one made with more recent data.

Appendix A

i. Contingency Tables for Association Tests

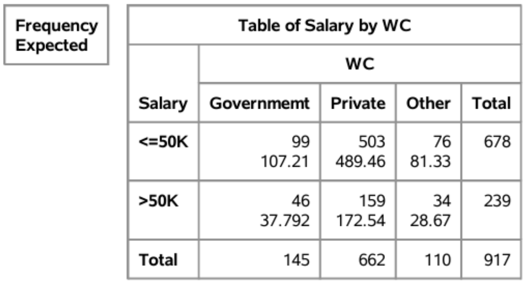
Figure 2.1



Frequency
Expected

Table of Salary by MS				
	MS			
Salary	No Longer Married	Married	Never Married	Total
<=50K	153 124.95	248 336.41	277 216.63	678
>50K	16 44.047	207 118.59	16 76.365	239
Total	169	455	293	917

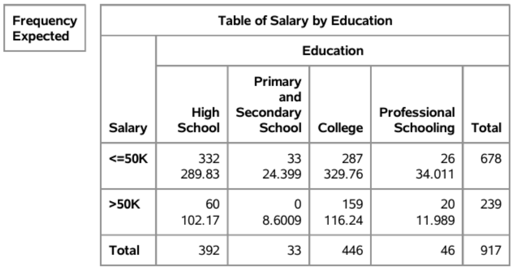
Figure 2.2



Frequency
Expected

Table of Salary by WC				
	WC			
Salary	Governmemt	Private	Other	Total
<=50K	99 107.21	503 489.46	76 81.33	678
>50K	46 37.792	159 172.54	34 28.67	239
Total	145	662	110	917

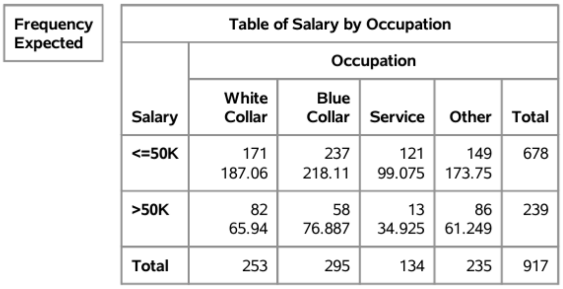
Figure 2.3



Frequency
Expected

Table of Salary by Education					
	Education				
Salary	High School	Primary and Secondary School	College	Professional Schooling	Total
<=50K	332 289.83	33 24.399	287 329.76	26 34.011	678
>50K	60 102.17	0 8.6009	159 116.24	20 11.989	239
Total	392	33	446	46	917

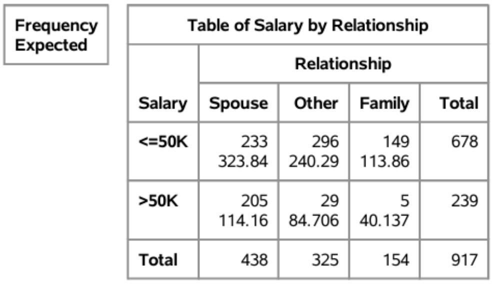
Figure 2.4



Frequency
Expected

Table of Salary by Occupation					
	Occupation				
Salary	White Collar	Blue Collar	Service	Other	Total
<=50K	171 187.06	237 218.11	121 99.075	149 173.75	678
>50K	82 65.94	58 76.887	13 34.925	86 61.249	239
Total	253	295	134	235	917

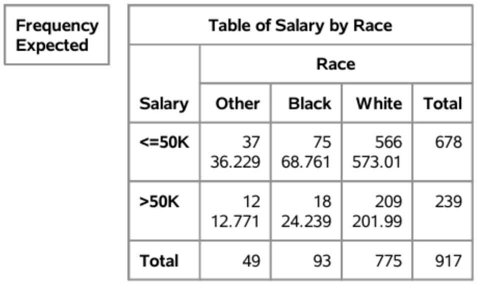
Figure 2.5



Frequency
Expected

Table of Salary by Relationship				
	Relationship			
Salary	Spouse	Other	Family	Total
<=50K	233 323.84	296 240.29	149 113.86	678
>50K	205 114.16	29 84.706	5 40.137	239
Total	438	325	154	917

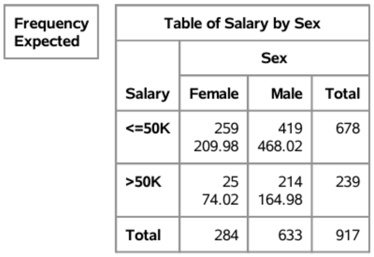
Figure 2.6



Frequency
Expected

Table of Salary by Race				
	Race			
Salary	Other	Black	White	Total
<=50K	37 36.229	75 68.761	566 573.01	678
>50K	12 12.771	18 24.239	209 201.99	239
Total	49	93	775	917

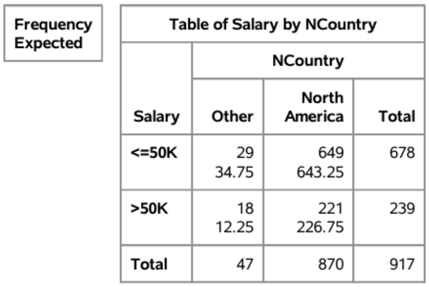
Figure 2.7



Frequency
Expected

Table of Salary by Sex			
	Sex		
Salary	Female	Male	Total
<=50K	259 209.98	419 468.02	678
>50K	25 74.02	214 164.98	239
Total	284	633	917

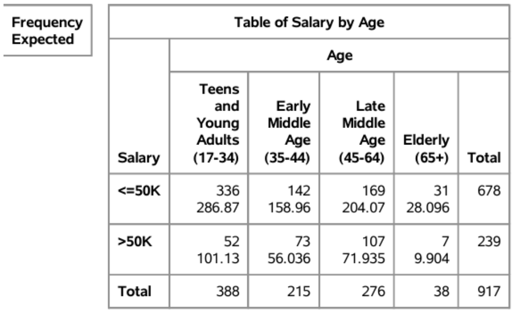
Figure 2.8



Frequency
Expected

Table of Salary by NCountry			
	NCountry		
Salary	Other	North America	Total
<=50K	29 34.75	649 643.25	678
>50K	18 12.25	221 226.75	239
Total	47	870	917

Figure 2.9

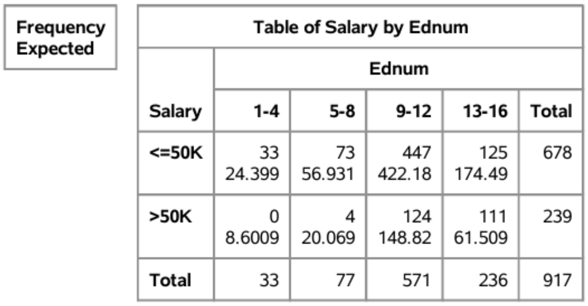


Frequency
Expected

Table of Salary by Age

	Age				
Salary	Teens and Young Adults (17-34)	Early Middle Age (35-44)	Late Middle Age (45-64)	Elderly (65+)	Total
<=50K	336 286.87	142 158.96	169 204.07	31 28.096	678
>50K	52 101.13	73 56.036	107 71.935	7 9.904	239
Total	388	215	276	38	917

Figure 2.10

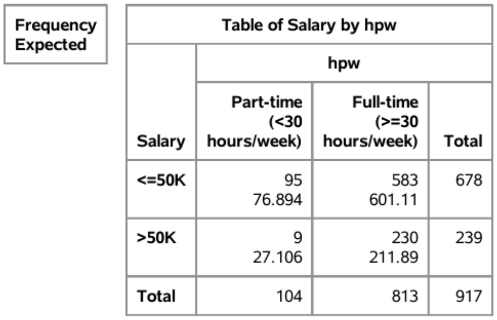


Frequency
Expected

Table of Salary by Ednum

	Ednum				
Salary	1-4	5-8	9-12	13-16	Total
<=50K	33 24.399	73 56.931	447 422.18	125 174.49	678
>50K	0 8.6009	4 20.069	124 148.82	111 61.509	239
Total	33	77	571	236	917

Figure 2.11



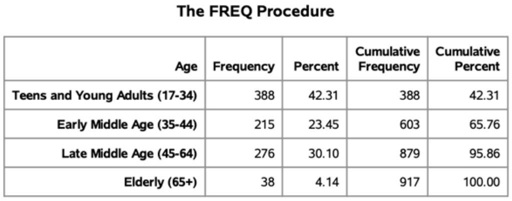
Frequency
Expected

Table of Salary by hpw

	hpw		
Salary	Part-time (<30 hours/week)	Full-time (>=30 hours/week)	Total
<=50K	95 76.894	583 601.11	678
>50K	9 27.106	230 211.89	239
Total	104	813	917

ii. Frequency Tables for Exploratory Data Analysis

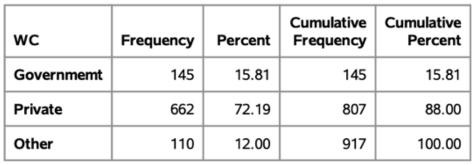
Figure 1.1



The FREQ Procedure

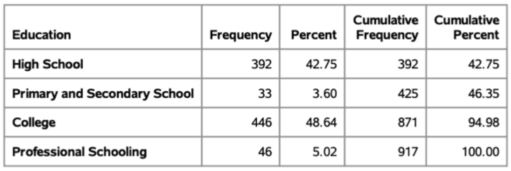
Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Teens and Young Adults (17-34)	388	42.31	388	42.31
Early Middle Age (35-44)	215	23.45	603	65.76
Late Middle Age (45-64)	276	30.10	879	95.86
Elderly (65+)	38	4.14	917	100.00

Figure 1.2



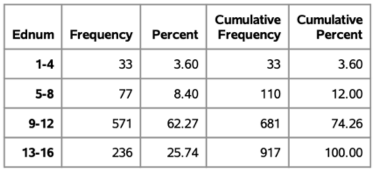
WC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Governmemt	145	15.81	145	15.81
Private	662	72.19	807	88.00
Other	110	12.00	917	100.00

Figure 1.3



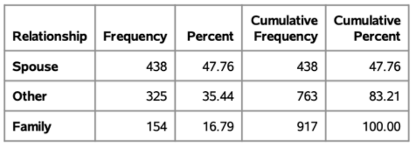
Education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High School	392	42.75	392	42.75
Primary and Secondary School	33	3.60	425	46.35
College	446	48.64	871	94.98
Professional Schooling	46	5.02	917	100.00

Figure 1.4



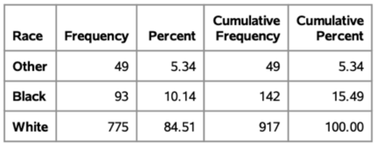
Ednum	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1-4	33	3.60	33	3.60
5-8	77	8.40	110	12.00
9-12	571	62.27	681	74.26
13-16	236	25.74	917	100.00

Figure 1.5



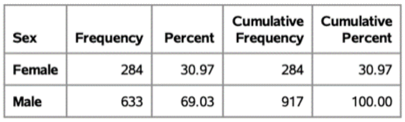
Relationship	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Spouse	438	47.76	438	47.76
Other	325	35.44	763	83.21
Family	154	16.79	917	100.00

Figure 1.6



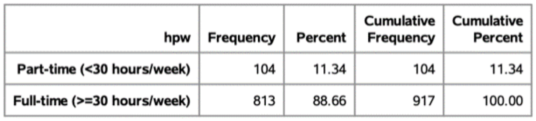
Race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Other	49	5.34	49	5.34
Black	93	10.14	142	15.49
White	775	84.51	917	100.00

Figure 1.7



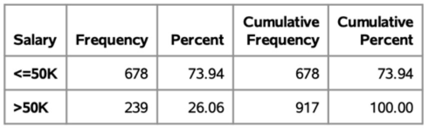
Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	284	30.97	284	30.97
Male	633	69.03	917	100.00

Figure 1.8



hpw	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Part-time (<30 hours/week)	104	11.34	104	11.34
Full-time (>=30 hours/week)	813	88.66	917	100.00

Figure 1.9



Salary	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<=50K	678	73.94	678	73.94
>50K	239	26.06	917	100.00