Deliverable 5:

Does average salary predict Glassdoor ratings?

Rebekah Sander

Research Question

Does average salary category predict the Glassdoor rating of the company for the job listing?

Research Variables

WHO	WHAT measurement is made on each		TYPE OF MEASURE
	Name of Variable	Question Asked	
Job Listing	Average Salary	What is the average salary of a job at this company?	Categorical Variable: levels = Under $81,000, $81,000 to $111,500, $111,500 and over
	Glassdoor Rating	What is the Glassdoor rating for this company?	Quantitative Variable Unit: Company Rating
One quantitative variable being tested among one categorical variable to see difference amongst levels.			
1. One-way ANOVA 2. Welch's Test on Raw Data	3. Kruskal Wallis 4. Welch's Test on Ranked Data		

SAS Code: Examining the Data

```
/*keeping and renaming*/
data work.data;
    set work.data_FULL (keep = Rating 'Avg Salary(K)'n);
    rename 'Avg Salary(K)'n ='Average Salary'n Rating='Glassdoor Rating'n;
run;

/* Check for and fix miscoding/missing values */
/* There is a -1 on Glassdoor rating*/
Proc Means data = work.data MAXDEC=2 n mean stddev median Qrange RANGE min Q1 Q3 max
    var 'Average Salary'n 'Glassdoor Rating'n; /*Quantitative variable*/
run;

PROC FREQ DATA=WORK.data;
TABLE 'Average Salary'n 'Glassdoor Rating'n;
run;

Proc Contents data=work.data varnum;
run;

data work.data_scientist;
    set work.data;
    if 'Glassdoor Rating'n = -1 then delete;
run;
```

Examining the Data

Variable	Label	N	Mean	Std Dev	Median	Quartile Range	Range	Minimum	Lower Quartile	Upper Quartile	Maximum
Average Salary	Avg Salary(K)	742	101.48	37.48	97.50	49.00	238.50	15.50	73.50	122.50	254.00
Glassdoor Rating	Rating	742	3.62	0.80	3.70	0.70	6.00	-1.00	3.30	4.00	5.00

Variable	Label	N	Mean	Std Dev	Median	Quartile Range	Range	Minimum	Lower Quartile	Upper Quartile	Maximum
Average Salary	Avg Salary(K)	731	100.96	37.14	96.50	49.50	238.50	15.50	73.00	122.50	254.00
Glassdoor Rating	Rating	731	3.69	0.57	3.70	0.70	3.10	1.90	3.30	4.00	5.00

Variables in Creation Order					
#	Variable	Type	Len	Format	Label
1	Glassdoor Rating	Num	8	BEST.	Rating
2	Average Salary	Num	8	BEST.	Avg Salary(K)

Creating Categories in Average Salary

```
/*Splitting avg salary into 3 groups*/
/*Getting the values for the 33rd and 66th percentile.*/
proc univariate data=work.data_scientist;
    var 'Average Salary'n;
    output out=work.salary_33_66
    pctlpts = 33, 66
    pctlpre = P_;
run;
```

```
Proc format;
Value $avgformat
'A'="Under $81,000"
'B'="$81,000 to $111,500"
'C'="$111,500 and over";
run;
```

```
data work.data_scientist;
set work.data_scientist;
format 'Average Salary Category'n avgformat.;
run;
```

```
data work.data_scientist;
set work.data_scientist;
length 'Average Salary Category'n $20;
    if 'Average Salary'n <81 then 'Average Salary Category'n='A';
    else if 81<= 'Average Salary'n <111.5 then 'Average Salary Category'n='B';
    else if 'Average Salary'n >=111.5 then 'Average Salary Category'n='C';
RUN;
```

SAS Code: Assessing Normality

```
PROC FREQ data=work.data_scientist;
tables 'Average Salary Category'n;
run;

/*QQ Plots and normality test*/
title 'Figures 1, 2, 3: QQ Plots for Glassdoor Ratings by Average Salary';
proc univariate data=work.data_scientist normaltest plots;
    var 'Glassdoor Rating'n; /*QUANTITATIVE*/
    class 'Average Salary Category'n; /*CATEGORICAL*/
title;
```

Assessing Normality

- H_0: The data came from a population where Glassdoor Ratings are normally distributed
- H_A: The data came from a population where Glassdoor Ratings are not normally distributed.

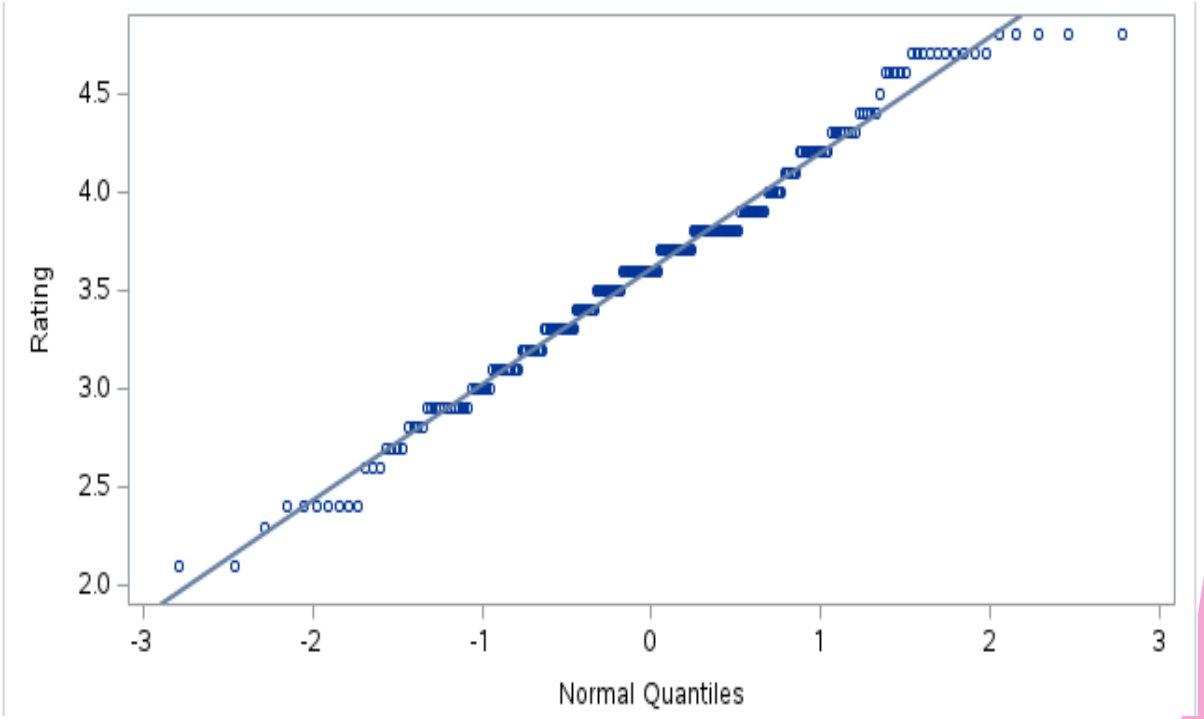
Average Salary Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Under $81,000	233	31.87	233	31.87
$81,000 to $111,500	246	33.65	479	65.53
$111,500 and over	252	34.47	731	100.00

- The sample sizes for all three categories are greater than 30. By the Central Limit Theorem, the $\bar{x}$ distribution is normal.

Assessing Normality—Under $81,000

- Normality Tests: All four of the tests for normality show p-values less than $\alpha = 0.05$. Thus, we have evidence to suggest the x distribution is not normal. (note: tests are sensitive to sample size)
- QQ Plot: The data follows the agreement line with little deviation. This supports that the x distribution is normal.

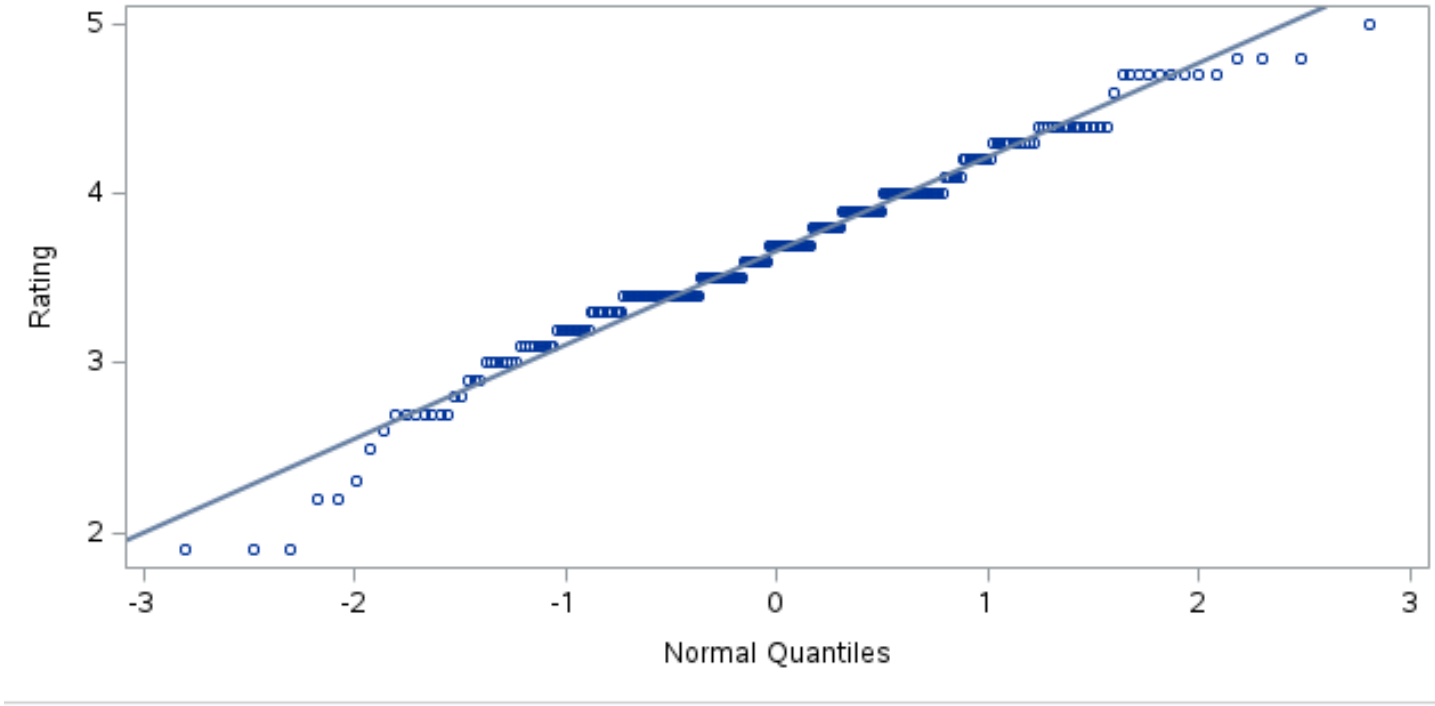
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.985075	Pr < W	0.0152
Kolmogorov-Smirnov	D	0.073992	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.131402	Pr > W-Sq	0.0435
Anderson-Darling	A-Sq	0.842198	Pr > A-Sq	0.0308



Assessing Normality—$81,000 to $111,500

- Normality Tests: All four of the tests for normality show p-values less than $\alpha = 0.05$. Thus, we have evidence to suggest the x distribution is not normal. (note: tests are sensitive to sample size)
- QQ Plot: The data follows the agreement line with little deviation. This supports that the x distribution is normal.

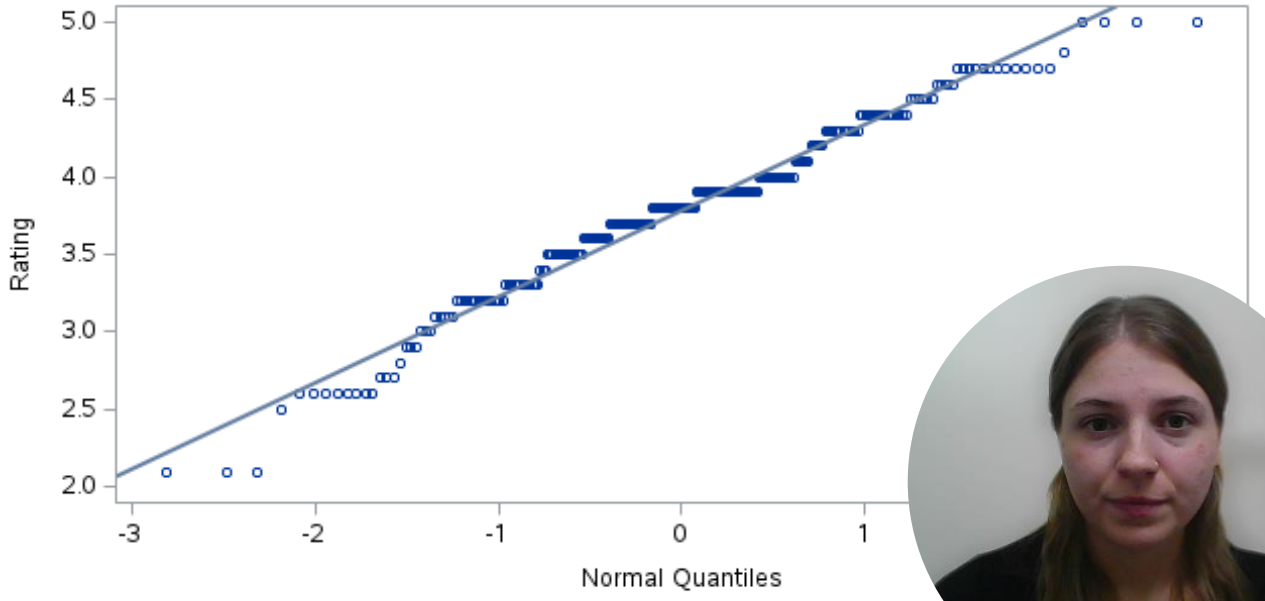
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.980135	Pr < W	0.0016
Kolmogorov-Smirnov	D	0.084038	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.173852	Pr > W-Sq	0.0120
Anderson-Darling	A-Sq	1.138455	Pr > A-Sq	0.0057



Assessing Normality —$111,500 and over

- Normality Tests: All four of the tests for normality show p-values less than $\alpha = 0.05$. Thus, we have evidence to suggest the x distribution is not normal. (note: tests are sensitive to sample size)
- QQ Plot: The data follows the agreement line with little deviation. This supports that the x distribution is normal.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.977738	Pr < W	0.0005
Kolmogorov-Smirnov	D	0.096386	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.320039	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.757951	Pr > A-Sq	<0.0050



Assessing Homogeneity

- H_0: Data is Homogeneous H_A: Data is Non-Homogeneous

```
/*A Rough Check for Homogeneity*/
TITLE "Table 1: Table to Compare Standard Deviations for Homogeneity";
PROC MEANS DATA = work.data_scientist mean stddev VAR maxdec=4;
      class 'Average Salary Category'n;
      VAR 'Glassdoor Rating'n;
RUN;
TITLE;
```

Table 1: Table to Compare Standard Deviations for Homogeneity

The MEANS Procedure

Analysis Variable : Glassdoor Rating Rating				
Average Salary Category	N Obs	Mean	Std Dev	Variance
Under $81,000	233	3.6112	0.5898	0.3479
$81,000 to $111,500	246	3.6659	0.5543	0.3072
$111,500 and over	252	3.7817	0.5567	0.3099

Assessing Homogeneity

- H_0: Data is Homogeneous

  H_A: Data is Non-Homogeneous
- To assess homogeneity, we will look at the ratio of the standard deviations.
- The ratio is less than 2. Thus, the standard deviations are close enough to use a test that requires homogeneity.
- Conclusion: Use one-way ANOVA.

Ratio of standard deviations:

$$= \frac{SD_{over\ \$111,500}}{SD_{under\ \$81,000}}$$

$$= \frac{3.7817}{3.6112}$$

$$= 1.0472$$

Choosing Hypothesis Test: One-Way ANOVA

- Since the data is normal and homogeneous, we will perform the one-way ANOVA.

$$H_0: \mu_{under\ \$81{,}000} = \mu_{\$81{,}000\ to\ \$111{,}500} = \mu_{\$111{,}500\ and\ above}$$

$$H_A: \mu_{under\ \$81{,}000} (=/\neq) \mu_{\$81{,}000\ to\ \$111{,}500}\ (=/\neq)\ \mu_{\$81{,}000\ to\ \$111{,}500}$$

$$\alpha = 0.05$$

- The null hypothesis is that all three average salary categories have the same company Glassdoor ratings for all job listings in each category.
- The alternative hypothesis is that at least one population of the average salary categories has a different Glassdoor rating from the other two populations of average salaries.
- The level of significance, $\alpha = 0.05$, tells us that 5% of the time we will conclude H_A *when* H_0 is actually true.

Performing One-Way ANOVA

- **F-Value:** The variance in average salary is 5.78 times the pooled variance of the Glassdoor rating within the average salary category.
- **P-Value:** There is a 0.32% chance of observing an F test statistic of 5.78 or more when all average salary categories have the same company ratings.
- **Conclusion:** Since 0.0032 is less than the significance level of 0.05, we are 95% confident that the average company rating for one or more average salary category is different than the other categories.

The ANOVA Procedure

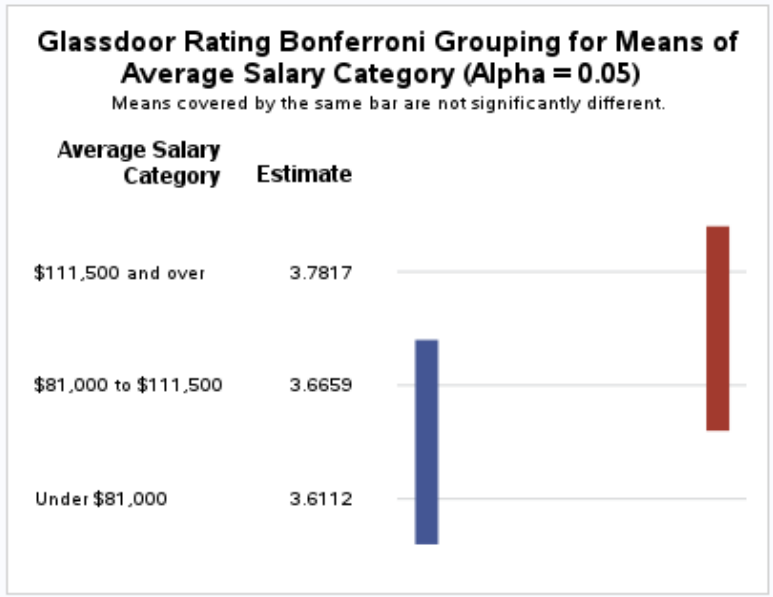
Dependent Variable: Glassdoor Rating Rating

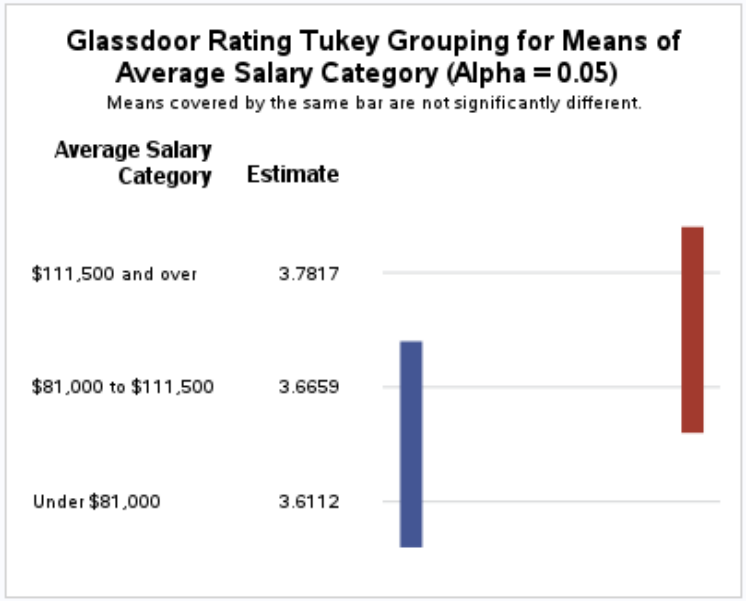
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.7109732	1.8554866	5.78	0.0032
Error	728	233.7601896	0.3210992		
Corrected Total	730	237.4711628			

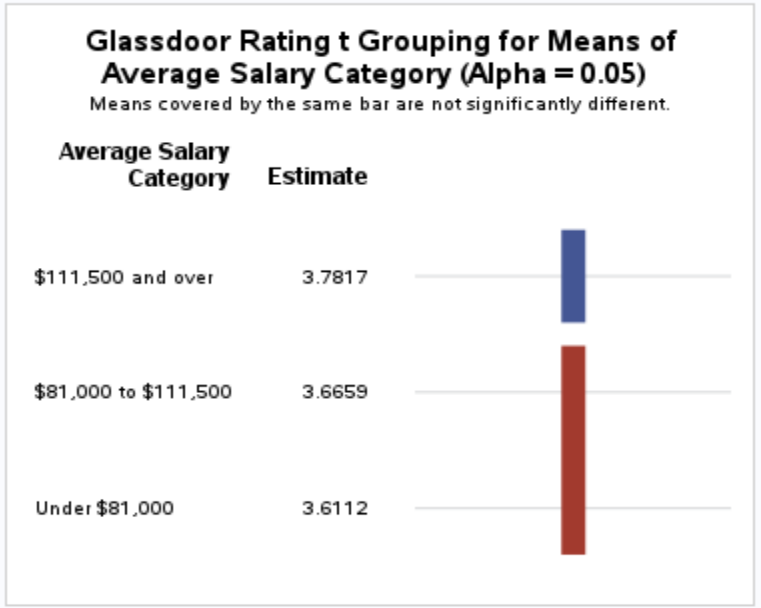
R-Square	Coeff Var	Root MSE	Glassdoor Rating Mean
0.015627	15.36331	0.566656	3.688372

```
proc anova data=work.data_scientist;
   class 'Average Salary Category'n;
   model 'Glassdoor Rating'n = 'Average Salary Category'
   means 'Average Salary Category'n / bon lines tukey li
run;
```

Post-hoc Tests



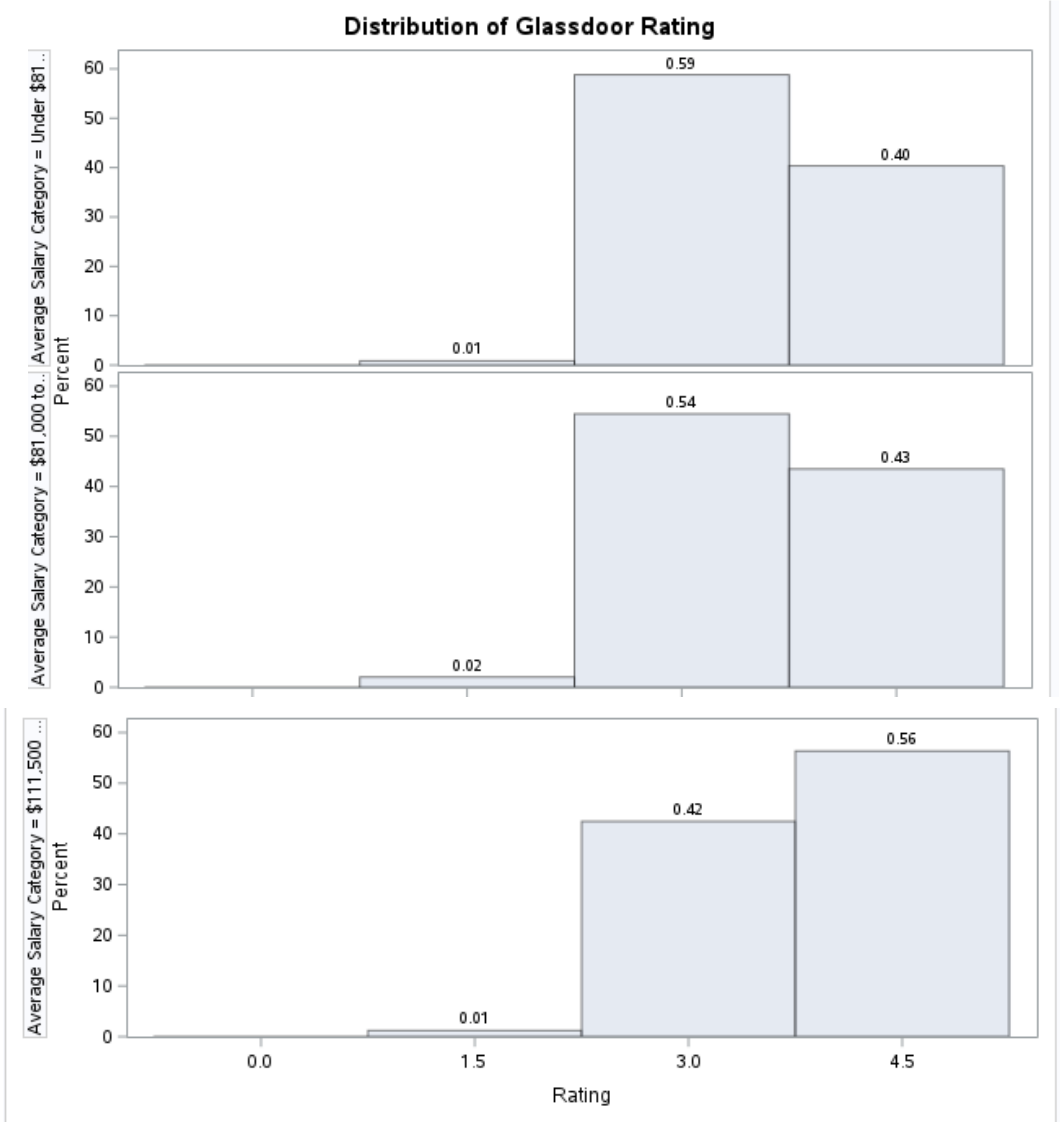




Bonferroni and Tukey: The average salary categories under $81,000 and $81,000 to $111,500 are not significantly different. $111,500 and over and $81,000 to $111,500 are not significantly different, $111,500 and over and under $81,000 are significantly different.

Fisher's Least Significant [illegible]:
The average salary ca
$81,000 and $81,000
not significantly diff
$111,500 and over is
different from the ot
categories.

Supporting Graphic: Histogram




```
/*Stratified Histogram*/
PROC UNIVARIATE DATA = work.data_scientist noprint;
  VAR 'Glassdoor Rating'n;
  CLASS 'Average Salary Category'n;
  HISTOGRAM/barlabel=proportion midpoints=0 to 5 by 1.5 ;
  TITLE1 height=16pt 'Figure 3:  Histogram of Casual Bike Rental Counts by Whether it is a Holiday';
  Title2 height=12pt 'Bin width is 300 rental counts';
RUN;
title;
```

Taking Action

- If a company wanted to improve their Glassdoor rating, they could raise salaries for their jobs available.
- If a person seeking a job at different companies wanted to see where they had the most potential to have a high salary, they may take Glassdoor ratings into consideration.

SAS Code Screen Recording