Data Analysis of the Pay Gap Between Genders

Matsik, Kelsey

Dataset Background and Description

Summary and Motivation for Choosing this Dataset

- I chose this dataset in order to examine the pay gap differences between males and females across several different professions. This data was collected through Glassdoor.
- **9 variables**
- **1,000 observations**

Description of Variables

1. *JobTitle*: Profession
2. *Gender*
3. *Age*
4. *PerfEval*: Performance Evaluation Score on a scale of 1-5 (1 being "Excellent" and 5 being "Poor")
5. *Degree*: Highest degree achieved
6. *Dept*: Department in which He/She Works
7. *Seniority*: Number of years worked
8. *BasePay*: Annual Base Pay in Dollars
9. *Bonus*: Annual Bonus Pay in Dollars
10. *TotalPay*: Annual Base Pay + Annual Bonus Pay in Dollars

Summary of Dataset

- **Categorical** Variables
  - JobTitle
  - Gender
  - Dept
  - Education
  - PerfEval
- **Numeric** Variables
  - Age
  - Seniority
  - BasePay
  - Bonus
  - TotalPay

Data Preparation and Summary

Data Cleaning

- Clean dataset
  - 0 rows containing nonresponse (NA) values
  - Good variable names
- PerfEval: Converted numeric values (1-5) to corresponding categorical value.
  - 5 → Excellent
  - 4 → Very Good
  - 3 → Good
  - 2 → Weak
  - 1 → Poor
- Education → Degree
  - High School → Diploma
  - College → Bachelors
- Addition of TotalPay (BasePay+Bonus)
- ***After*** Data Cleaning:
  - **10 variables** (5 categorical, 5 numeric)
  - **1,000 observations**

Summary Measures of Numeric Variables

Female Total Pay

Min.	: 40828
1st Qu.	: 80866
Median	: 96571
Mean	: 96417
3rd Qu.	:112660
Max.	:168968

Male Total Pay

Min.	: 41030
1st Qu.	: 87792
Median	:105100
Mean	:104919
3rd Qu.	:121617
Max.	:184010

Male Age

Min.	:18.00
1st Qu.	:28.00
Median	:40.00
Mean	:41.01
3rd Qu.	:55.00
Max.	:65.00

Female Age

Min.	:18.00
1st Qu.	:30.00
Median	:42.00
Mean	:41.83
3rd Qu.	:54.00
Max.	:65.00

Female Average Seniority
3.01

Male Average Seniority
2.93

Summary Measures of Categorical Variables

<u>Total Counts of Males and Females</u>

Female	Male
468	532

<u>Distribution of Female Professions</u>

Data Scientist	Driver	Financial Analyst	Graphic Designer
53	46	49	48
IT	Manager	Marketing Associate	Sales Associate
50	18	107	43
Software Engineer	Warehouse Associate		
8	46		

<u>Distribution of Male Professions</u>

Data Scientist	Driver	Financial Analyst	Graphic Designer
54	45	58	50
IT	Manager	Marketing Associate	Sales Associate
46	72	11	51
Software Engineer	Warehouse Associate		
101	44		

Summary Measures of Categorical Variables (ctd.)

What departments do males work in?

Administration	Engineering	Management	Operations	Sales
98	103	111	114	106

What departments do females work in?

Administration	Engineering	Management	Operations	Sales
95	89	87	96	101

Highest Degree Achieved - Female

Bachelors	Diploma	Masters	PhD
123	132	107	106

Highest Degree Achieved - Male

Bachelors	Diploma	Masters	PhD
118	133	149	132

Summary Measures of Categorical Variables (ctd.)

Female Performance Evaluation Scores

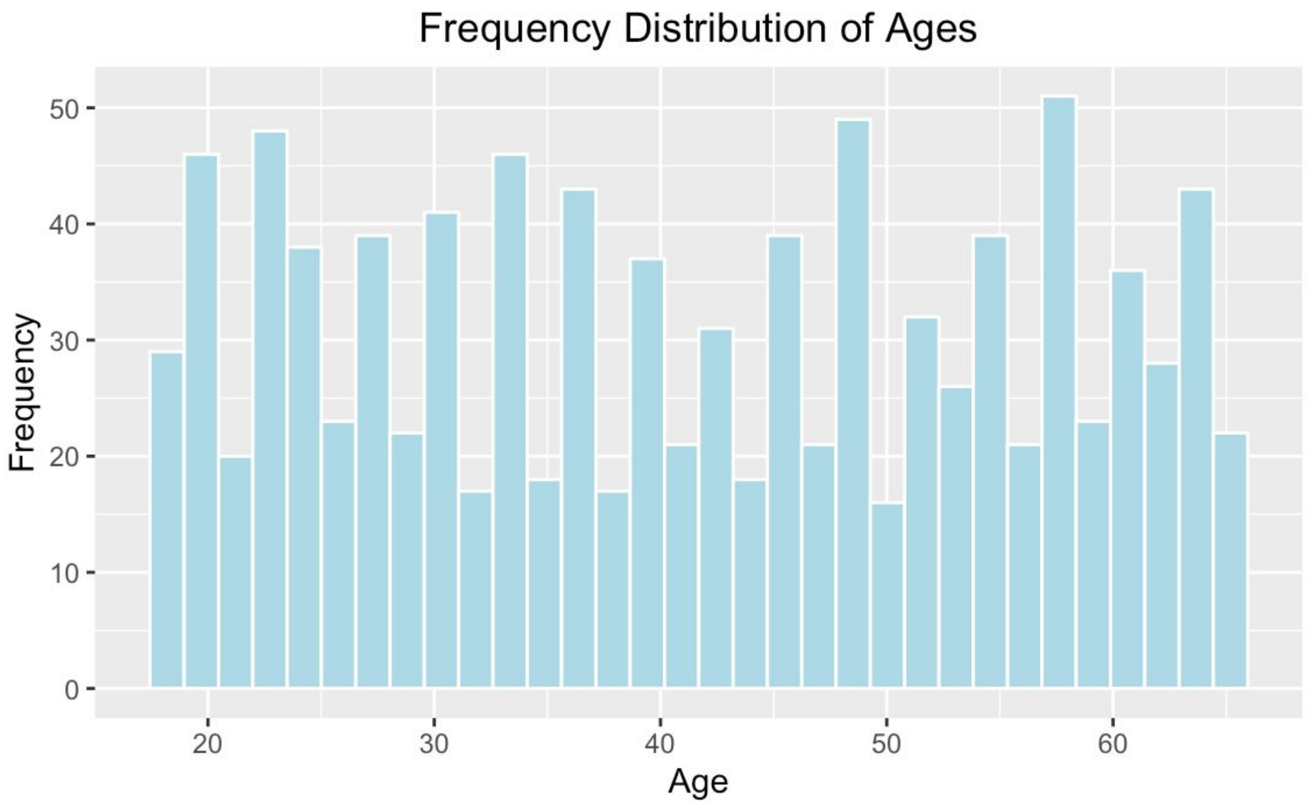
Excellent	Good	Poor	Very Good	Weak
88	88	106	96	90

Male Performance Evaluation Scores

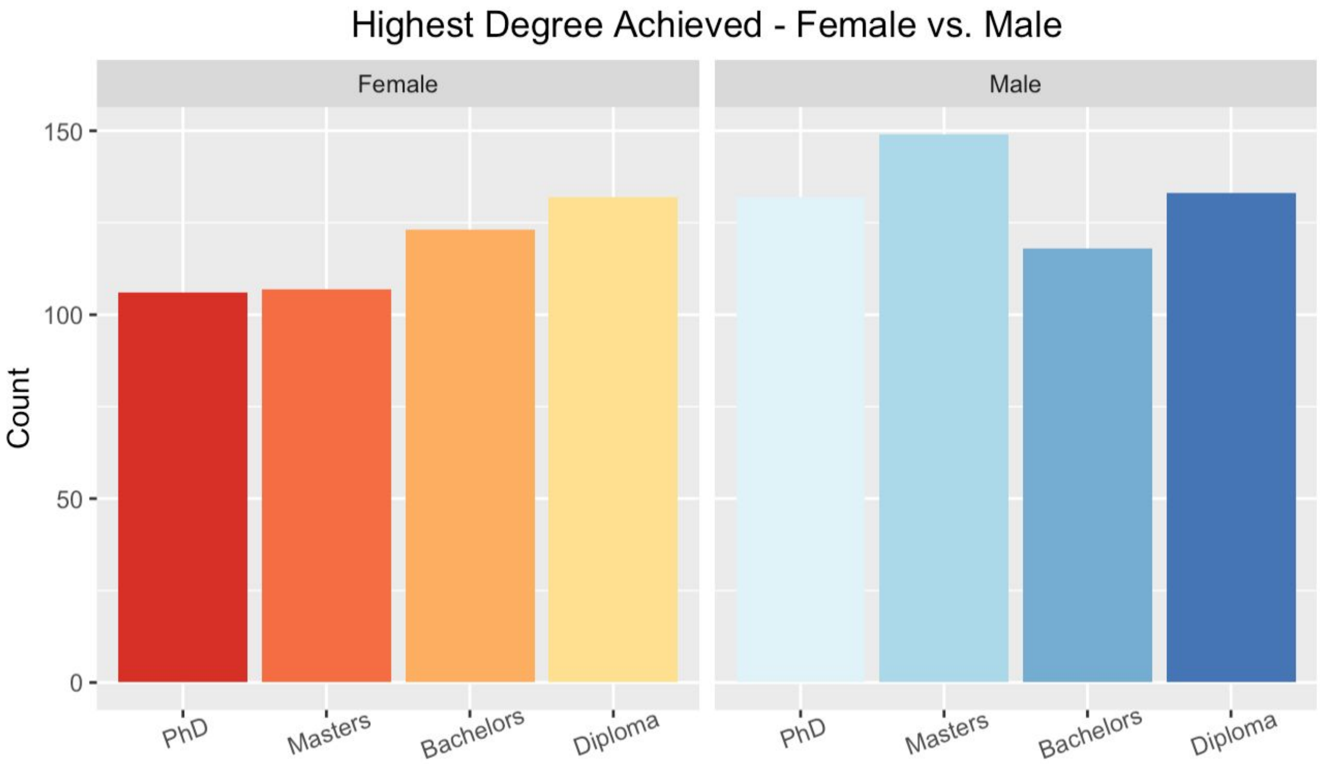
Excellent	Good	Poor	Very Good	Weak
121	106	92	111	102

Data Visualization

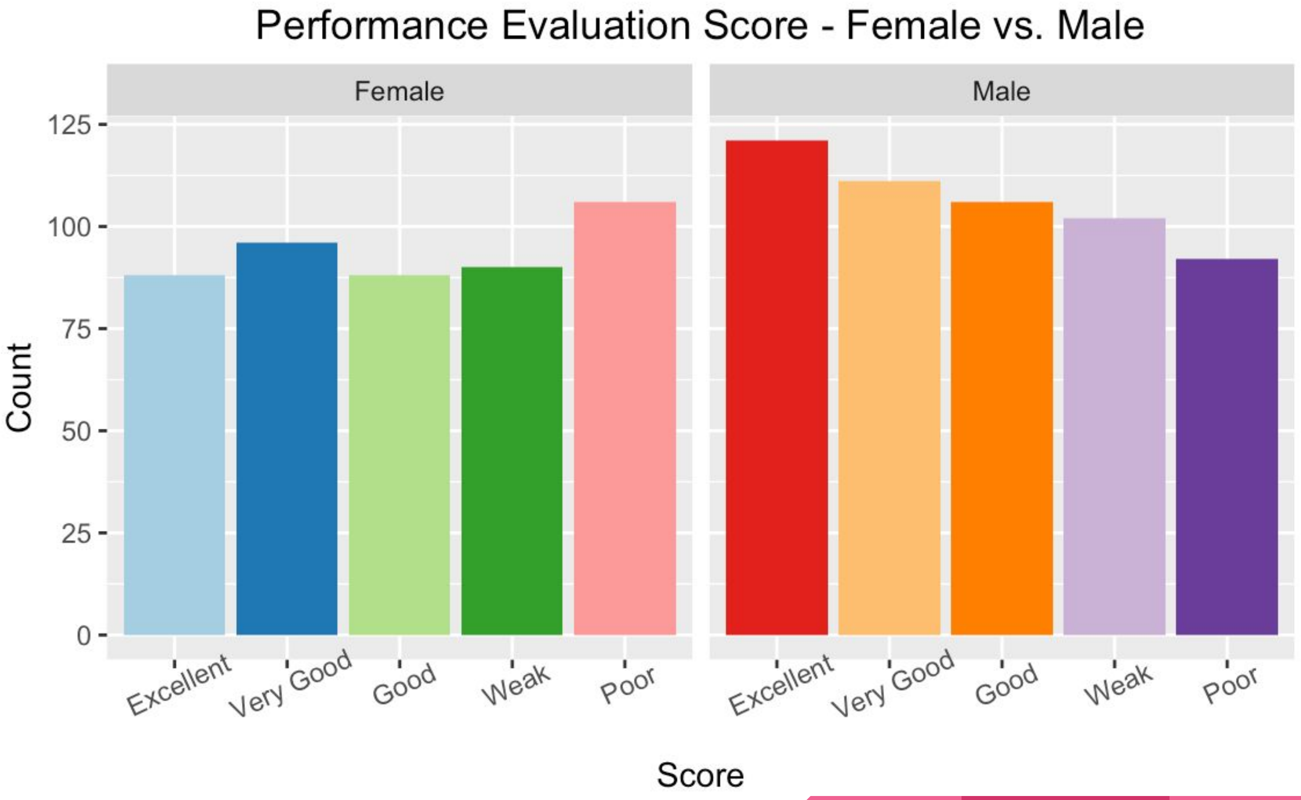
Histogram - What is the distribution of ages in the dataset?



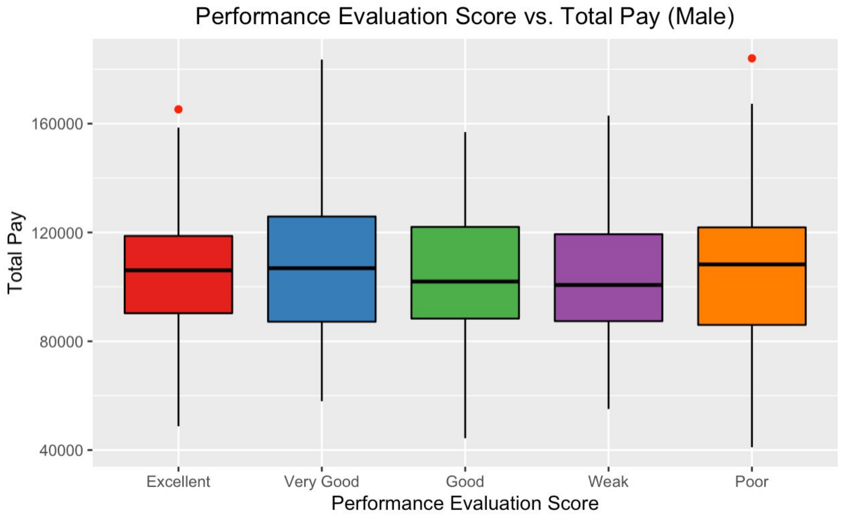
Bar Graph - What is the highest degree achieved between males and females?

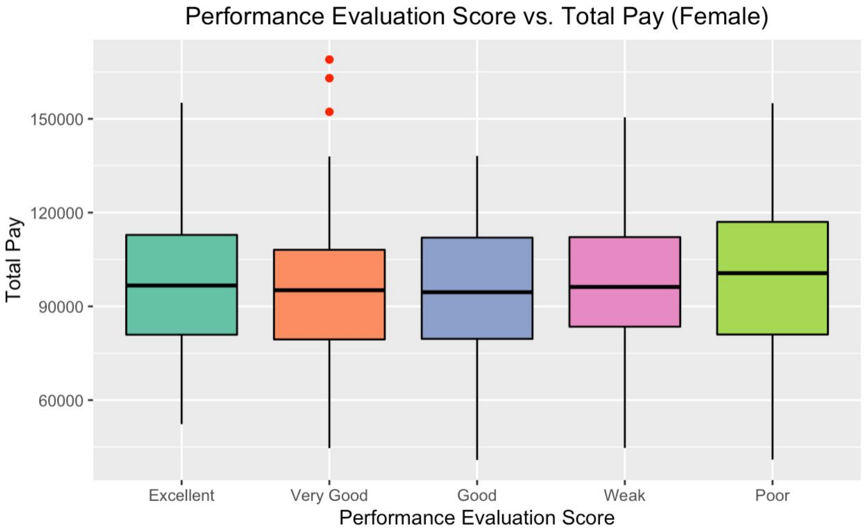


Bar Graph - How were males and females ranked on their performance evaluations?

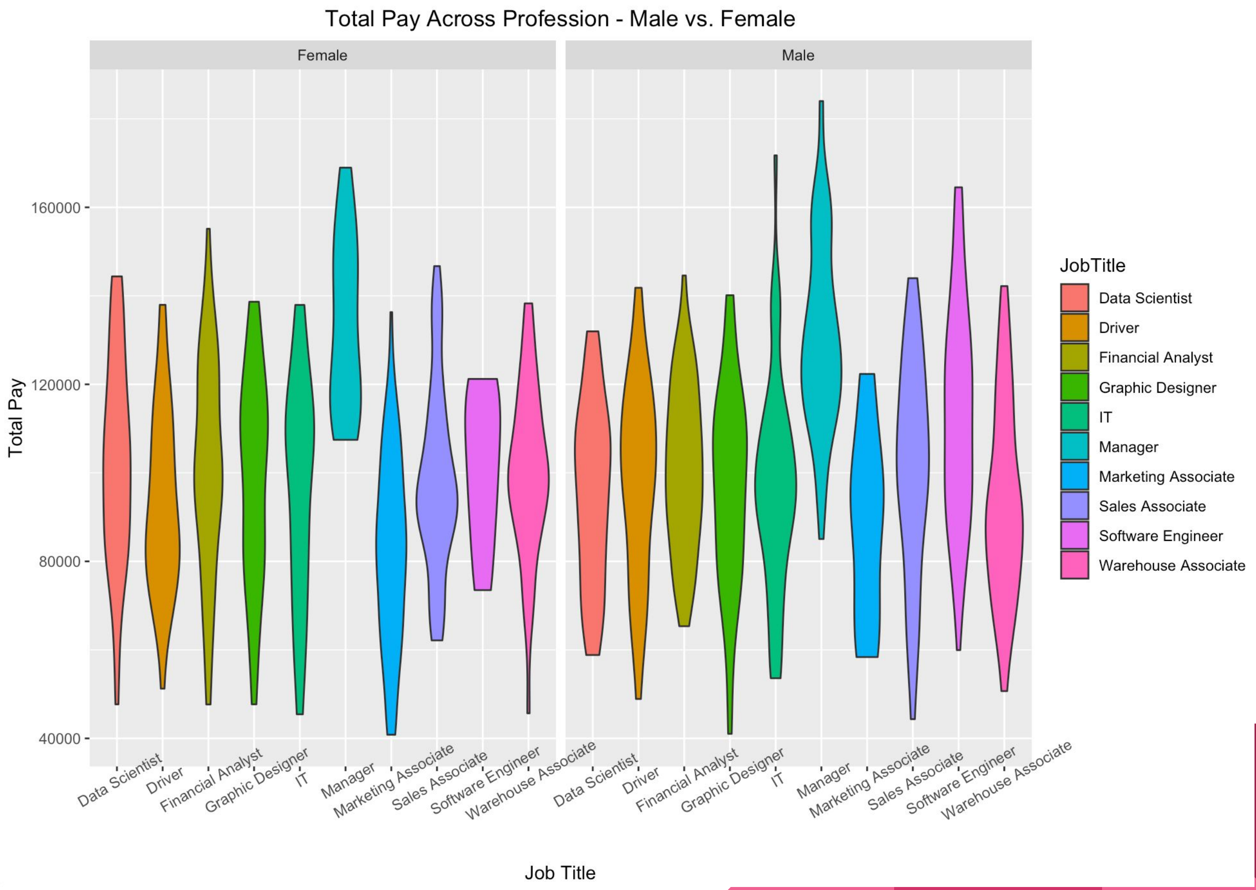


Side-by-Side Box Plot - What is the difference in total pay based on performance evaluation scores between males and females?

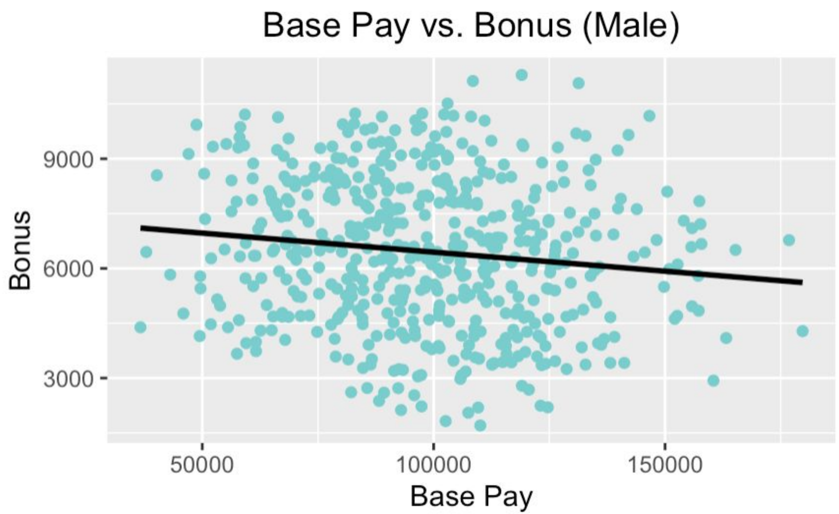


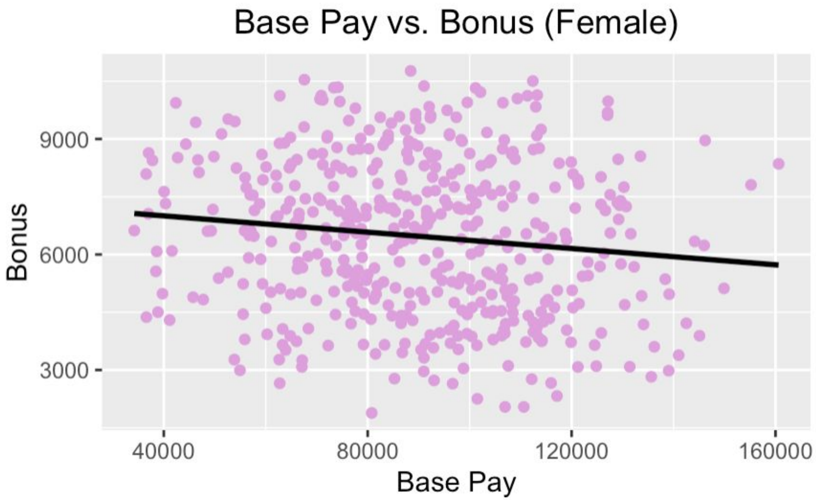


Violin Plot - How does total pay vary across profession between males and females?



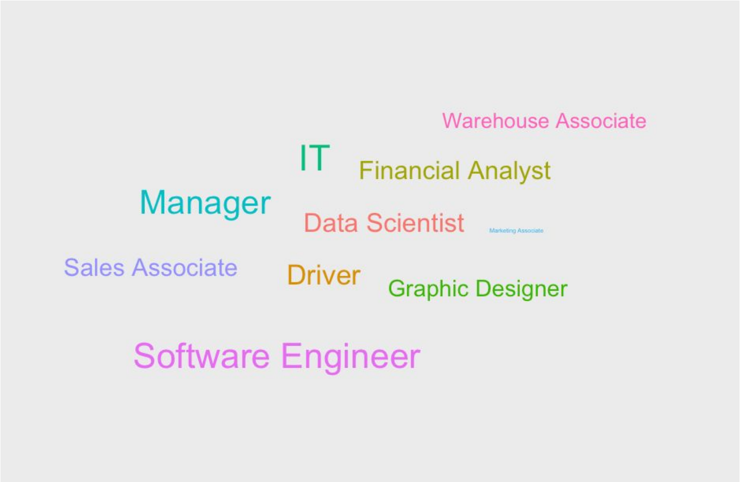
Scatterplot - What is the relationship between total pay and age between males and females?



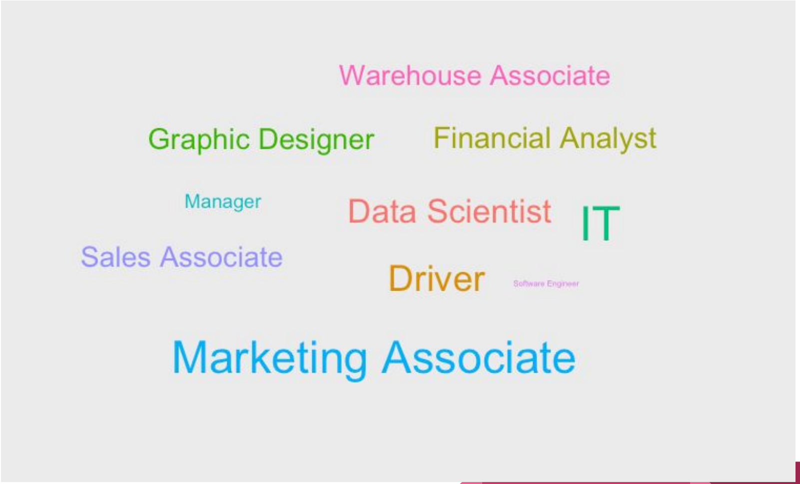


Word Cloud - What are the most common professions among males and females?

Males:



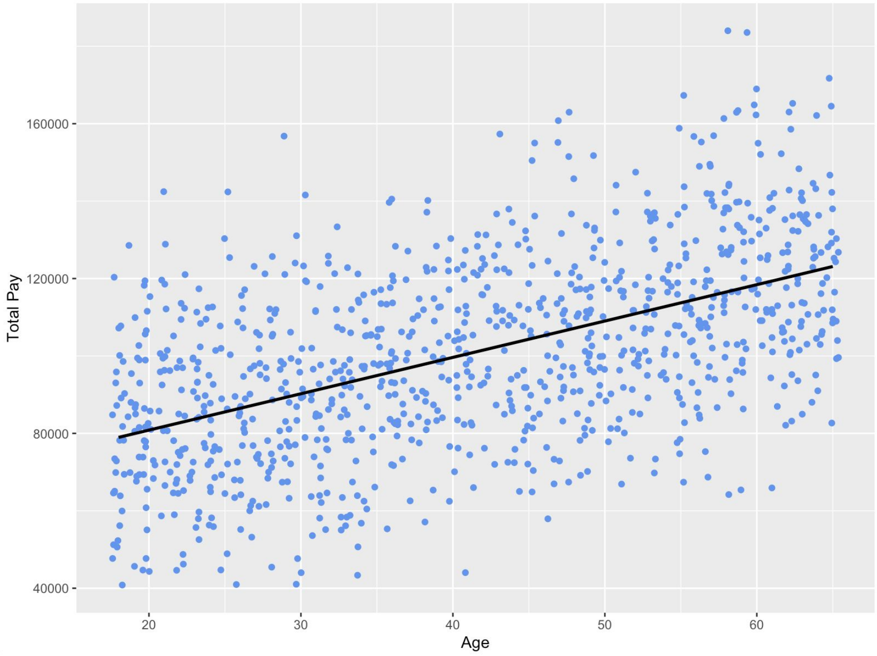
Females:



Regression

Simple Linear Regression - Predicting Total Pay from Age

Scatterplot - Age vs. Total Pay



Building the SLR Model

```
Coefficients:
(Intercept)          Age
    62061.4        939.3
```

TotalPay = 62,061.40 + 939.3**Age*

Interpretation in Context:

- Slope⇒ For every 1 year older the worker is, his/her total pay will increase by approximately $939.30.
- Y-intercept⇒ When the worker is 0 years old, they will earn $62,061.40 as their total pay.

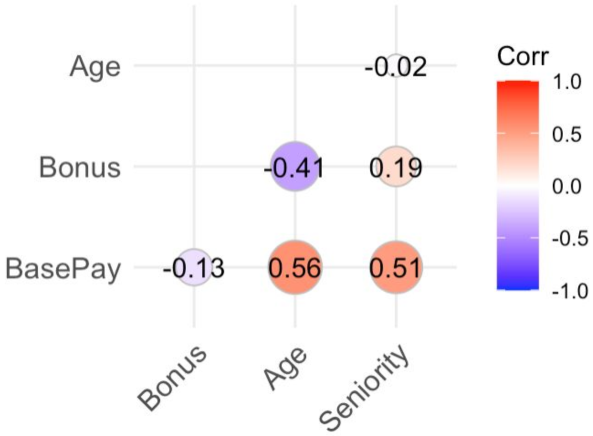
Simple Linear Regression - Predicting Total Pay from Age (ctd.)

Making a Prediction:

- ➢ Age (predictor) = 43
- ➢ Total Pay = $102,449.20
- ➢ If a worker is 43 years old, our model predicts that they will earn $102,449.20 as their total pay.

Multiple Linear Regression - Predicting Bonus from Age, Base Pay, and Seniority

Correlation Matrix



Building the MLR Model

1. "Kitchen sink" model

```
Coefficients:
(Intercept)      BasePay          Age    Seniority
  8.025e+03    1.208e-03   -5.877e+01    2.560e+02
```

2. Check for multicollinearity.

```
 BasePay       Age Seniority
2.441507  1.804566  1.669267
```

   a. All values <5 ⇒ multicollinearity is not an issue.

3. "Kitchen sink" model = Best model

MLR Model:

Bonus = 8025 + 0.0012**BasePay* – 58.77**Age* + 256**Seniority*

Multiple Linear Regression - Predicting Bonus from Age, Base Pay, and Seniority (ctd.)

Making a Prediction

- ➢ Predictor Variables
  - ○ Age = 37
  - ○ BasePay = $89,792
  - ○ Seniority = 2.75
- ➢ Response Variable
  - ○ Bonus = $6,663.08
- ➢ If a worker is 37 years old, earns $89,792 as their base pay, and has 2.75 years of experience, our model predicts that they will earn approximately a $6,663.08 bonus.

Logistic Regression - Predicting Gender from Performance Evaluation Score, Base Pay and Bonus

Using PayGap.Orig…
I used the original dataset (before cleaning) for my logistic regression analysis in order to use PerfEval (performance evaluation score) as a numeric variable.

Building the LR Model

1. Predictor Variables:
   a. BasePay
   b. Bonus
   c. PerfEval
2. Coefficients:

```
(Intercept)        Bonus      BasePay     PerfEval
 -9.152e-01   -1.818e-04    1.297e-05    3.294e-01
```

3. Check for multicollinearity.

```
   Bonus  BasePay PerfEval
3.896111 1.025748 3.845815
```

4. "Kitchen sink" model = best model

Predicting Gender

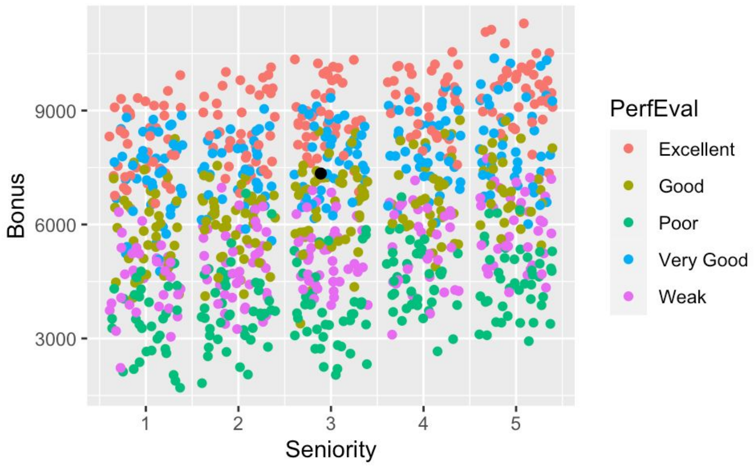
1. Sample Data:
   a. BasePay=$109,118
   b. Bonus=$10,578
   c. PerfEval=4
2. Prediction (Male or Female) ⇒ Female

Classification

KNN - Predicting Performance Evaluation Score Based on Bonus and Seniority

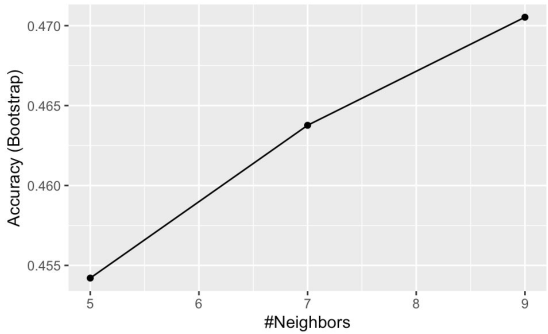
Using KNN, what would be the performance evaluation score of a worker who has 2.89 years of experience and earns a $7,342 bonus?

Visualize It: Scatterplot - Seniority vs. Bonus



In order to maximize accuracy, the model uses 9 neighbors to predict performance evaluation score (PerfEval).

K=9



Performance Evaulation Score Prediction: “Very Good”

Thank You!

References

https://www.datanovia.com/en/blog/the-a-z-of-rcolorbrewer-palette/

https://www.r-bloggers.com/2013/09/how-to-expand-color-palette-with-ggplot-and-rcolorbrewer/

https://ggplot2.tidyverse.org/reference/#plot-basics

https://r-graph-gallery.com/colors.html

https://www.statology.org/glm-fit-algorithm-did-not-converge/

https://www.performyard.com/articles/performance-review-ratings-scales-examples