Introduction

In this case study, I used the data collected through an anonymous web form on Fishbowl to evaluate salary insights across different Companies regarding Work experience, education level, ethnicity and Work satisfaction.
I tried to clean, process and Theorize the data available to me to determine trends and patterns, so that I can share some of the insights with the rest.

Because talking about how much or how little money you make feels taboo, and it shouldn't. Knowledge is power and Glassdoor info is hit or miss. Wouldn't it be great to know what your peers make so you can use that to leverage a raise? Or if your company makes a market adjustment yet you don't see the data, wouldn't it be great to know how accurate it is or isn't? So, let's share what we make and any relevant info to help each other learn our worth.

Deliverables

- Clear statement of the Objective
- Description of the data utilized
- Documentation of any data cleaning
- Supporting Visualizations and key findings

Objective

The Key task of this case study is to process and evaluate how salary distribution among working Individuals varies in different organizations.

Prepare

Download data and store it appropriately

Data has been downloaded from RealCPGSalaries. Local copies have been stored securely on Google Drive and GitHub

Process

Check the data for errors
The data is in CSV (comma-separated values) format, and there are a total of 24 columns.

Cleaning and filtering was done in excel by removing Duplicates and by using many other data sorting and cleaning tools.

Install and load necessary packages

```
> install.packages("tidyverse")
  install.packages("lubridate")
  install.packages("ggplot2")
  install.packages("dplyr")
  install.packages("geosphere")
```

Import data to R studio

```
> data <- read.csv("/cloud/project/Book2.csv")
> view(data)
>
```

Converted Salary and work_experience from chr to Numeric in order to run calculations.

Descriptive analysis on base_salary

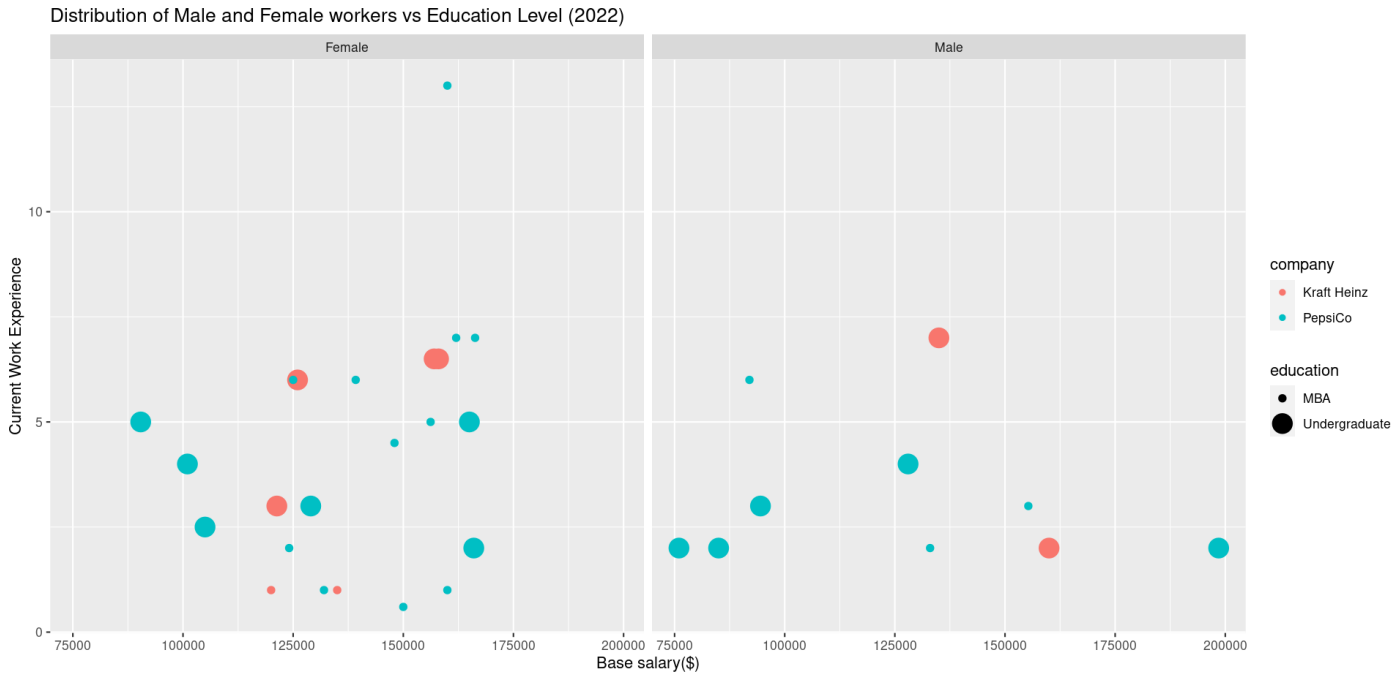
Mean = straight average for base_salary
DIfference = Numeric gap between the highest and the Lowest pay.
Lowest= Minimum pay
Highest = Maximum pay

```
> data %>%
+ drop_na(gender) %>%
+ group_by(gender) %>%
+ summarise(Lowest = min(base_salary), Highest = max(base_salary), Ave
rage = mean(base_salary), Difference = max(base_salary), min(base_sala
ry)) %>%
+ arrange(Average)
# A tibble: 2 × 6
  gender Lowest Highest Average Difference `min(base_salary)`
  <chr>   <int>   <int>   <dbl>      <int>               <int>
1 Male    76000  198500 127252.     198500               76000
2 Female  80000  175000 128041.     175000               80000
>
```

From the above Table we can observe that males tend to have the lowest and highest base salaries
Whereas Females tend to have a Higher Average Base salary.

Let’s see the Education Levels between males and Females

```
> data %>%
+ filter(company %in% c("PepsiCo","Kraft Heinz"))%>%
+ ggplot(aes(x = base_salary, y = work_experience, size = education, c
olor = company))+
+ geom_point()+
+ facet_wrap(~gender)+
+ labs(title = "Distribution of Male and Female workers vs Education L
evel (2022)", x = "Base salary($)", y = "Current Work Experience")
Warning message:
Using size for a discrete variable is not advised.
>
```



From the above graph it is observed that most of the males and females working in PepsiCo and Kraft Heinz have under 6 years of experience with Undergraduate education. Despite those factors they seem to have the same Pay Rate as the individuals with MBA Background.

Let's visualize Companies Vs Yearly Bonuses

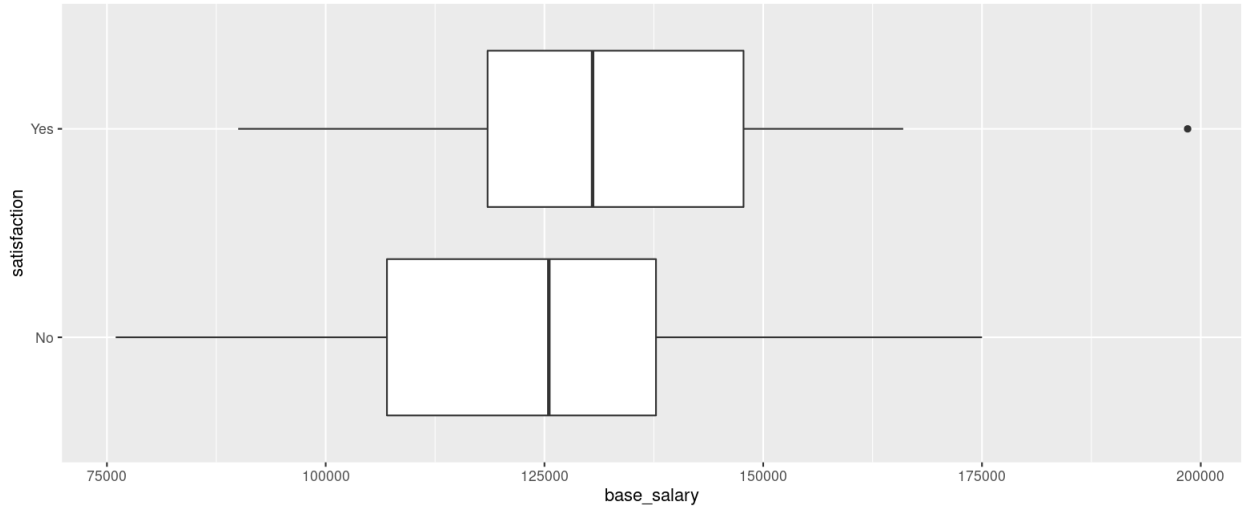
```
> data %>%
+ filter(company %in% c("Pepsico", "Kraft Heinz", "Mondelez", "Unileve
r", "SC Johnson")) %>%
+ ggplot(aes(x= base_salary, y = bonus, size = work_experience, color
= company))+
+ geom_point()+
+ labs(title = "Salary Distribution among Companies vs work Experienc
e", x = Base salary, y = Percentage of yearly Bonus))
```

From above I can tell Kraft Heinz employees are rich!

Let's Visualize the Satisfaction rate of all Individuals

Due to Numerios empty responses in satisfaction col I had to use drop_na to filter out empty spaces for a consistent representation.

```
> data %>%
+ drop_na(satisfaction) %>%
+ ggplot(aes(satisfaction, base_salary)) +
+ geom_boxplot() +
+ coord_flip()
>
```

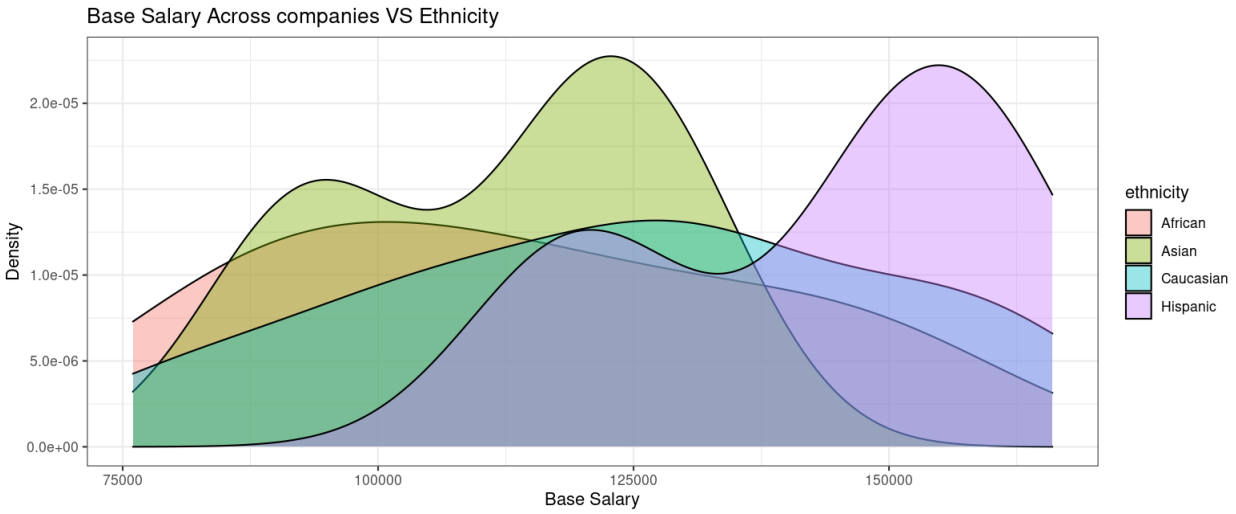


From above we see mostly people are not satisfied with their pay. From the data available I can say that the sweet stop for satisfaction lies between 125-130k.

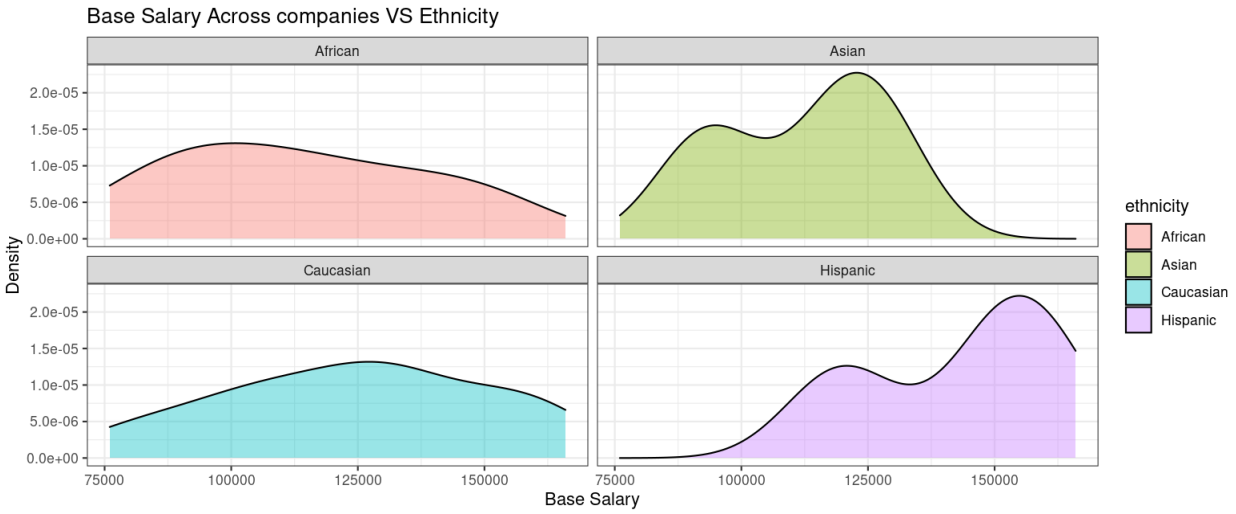
Let's compare base salary across different ethnicities

Again I was getting data shown for NA participants so had to filter those out.

```
> data %>%
+ drop_na(ethnicity) %>%
+ ggplot(aes(base_salary, fill = ethnicity)) +
+ geom_density(alpha = 0.4) +
+ theme_bw()+
+ facet_wrap(~ethnicity)+
+ labs(title = "Base Salary Across companies VS Ethnicity", x = "Base
Salary", y = "Density")
>
```



Here the data is really brought to life but still it is hard to properly make a statement due to high data traffic, so lets visualize them Individually.



Looks better, the base pay highly varies across different ethnicities with Hispanics being the top earners, while the others are pretty consistent across the graph.

Few Key Metrics

- Most of the Individuals working in these companies have Undergraduate Education.
- Kraft Heinz employees being the top earners despite having similar or even less experience as compare to their counterparts from other companies
- Satisfaction pay lies under 130k for the majority, while Females being the Top Average Earners in the given dataset.

Resources

Fishbowl

Dataset, Survey and GitHub