First\_Project

Robert Tarus

2022-05-06

# Google Data Analyst Professional Course Capstone Project.

## Human Resource Analytics.

### Scenario.

You are a junior data analyst working for a business intelligence consultant. You have been at your job for six months, and your boss feels you are ready for more responsibility. He has asked you to lead a project for a brand new client — this will involve everything from defining the business task all the way through presenting your data-driven recommendations. You will choose the topic, ask the right questions, identify a fresh dataset and ensure its integrity, conduct analysis, create compelling data visualizations, and prepare a presentation.

### Ask Phase.

I will be analyzing a dataset about a company with employees working from different parts of the world. The dataset is sourced from Alfie Danish. I will be analyzing the salary ranges across the countries and also the salary scale across gender and department. The main objective is to find out whether there is a disparity in the salary scale across the board.

### Prepare Phase.

The data is located in <https://www.kaggle.com/datasets/alfiedanish/hr-employee-dataset-1>. It is organized into 8 columns and 1438 rows. There is an extra column for salary, some of the columns haven’t been formatted, a city name by the name SÃ£o Paulo will need to be rectified. There are five instances of the word “Remote SumUp Limited” in the city column. There is one instance of the word “Remote Sumup Services” in the city column.

### Process Phase.

I will be using Excel and R to clean the data, SQL to explore and filter the data and R for visualizing the data.

## Setting up my environment

Notes: Loading the ‘tidyverse’, ‘dplyr’, ‘tidyr’, ‘ggplot2’, ‘ggstatsplot’, ‘scales’, ‘stringr’, ‘lubridate’ packages.

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.6 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readxl)  
library(ggstatsplot) #Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

## Registered S3 method overwritten by 'parameters':  
## method from   
## format.parameters\_distribution datawizard

## You can cite this package as:  
## Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.  
## Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(ggplot2)  
library(ggfortify)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

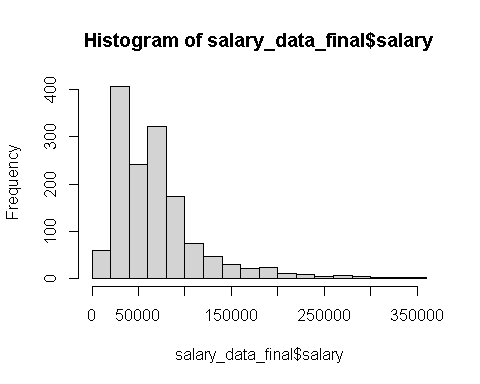
### Importing Dataset

library(readxl)  
salary\_data\_final <- read\_excel("salary\_data\_final.xlsx",   
 col\_types = c("text", "numeric", "text",   
 "text", "text", "date", "text", "text"))  
View(salary\_data\_final)

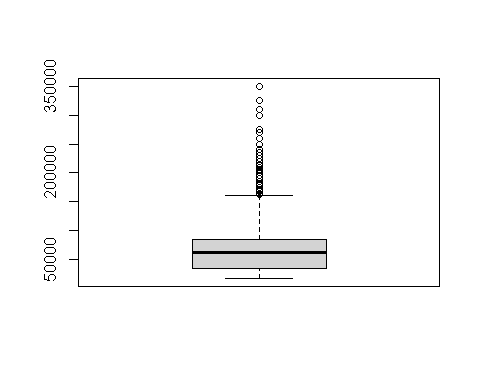
### Visualizations

#### Visual Exploratory Analysis.

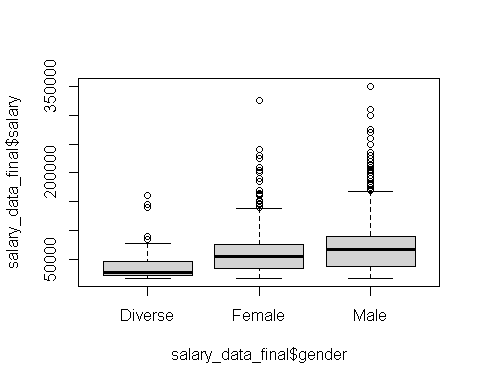
hist(salary\_data\_final$salary)



boxplot(salary\_data\_final$salary)



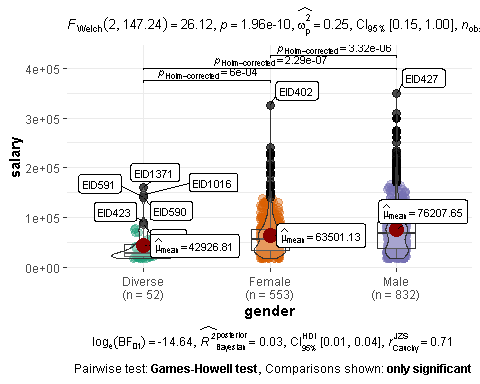
boxplot(salary\_data\_final$salary~salary\_data\_final$gender, data=salary\_data\_final)



### Labeling Outliers

ggbetweenstats(data = salary\_data\_final,  
 x = gender,  
 y = salary,  
 outlier.tagging = TRUE,  
 outlier.label = employee\_id)

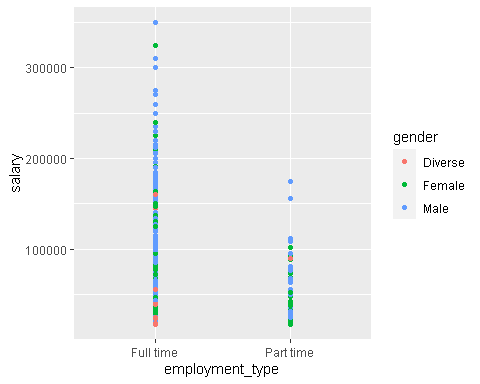
## Warning: ggrepel: 87 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps



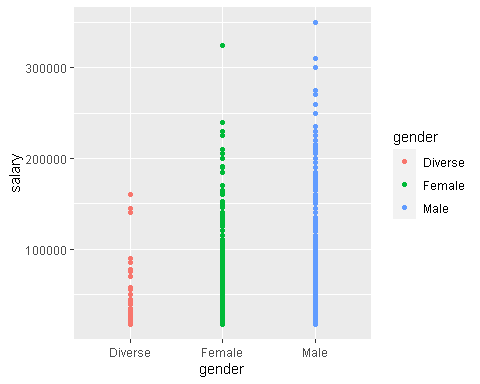
## Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

### Scatter plot

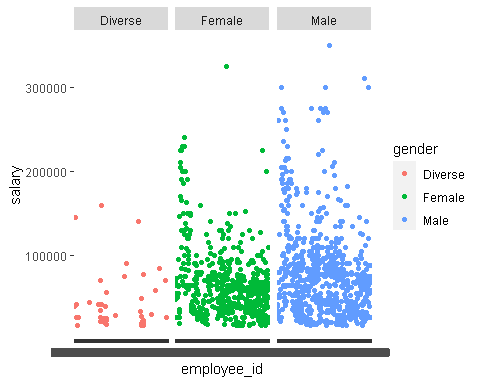
## Comparison between full time salary and part time salary  
ggplot(salary\_data\_final, aes(x=employment\_type, y=salary)) +  
 geom\_point(aes(col=gender)) + scale\_y\_continuous(labels=function(n){format(n, scientific = FALSE)})



# salary vs gender scatter plot  
ggplot(salary\_data\_final, aes(x=gender, y=salary)) +  
 geom\_point(mapping=aes(color=gender)) + scale\_y\_continuous(labels=function(n){format(n, scientific = FALSE)})

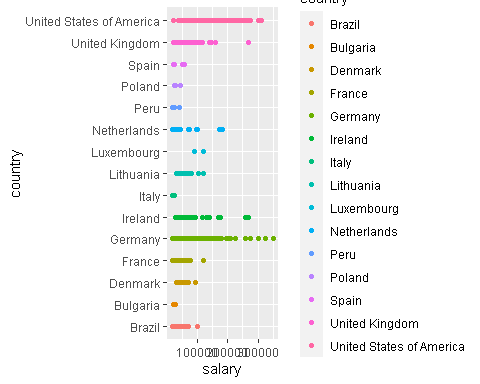


# salary vs employee\_id across genders  
ggplot(salary\_data\_final, aes(x=employee\_id, y=salary)) +  
 geom\_point(aes(color=gender)) + facet\_wrap(~gender) +   
 scale\_y\_continuous(labels=function(n){format(n, scientific = FALSE)})

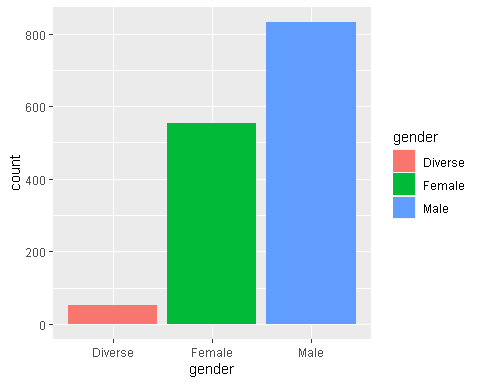


## Salary distribution across the different countries  
ggplot(salary\_data\_final, aes(x=salary, y=country)) +  
 geom\_point(aes(col=country)) + geom\_smooth(method="loess", se=F) +  
 scale\_x\_continuous(labels=function(n){format(n, scientific = FALSE)})

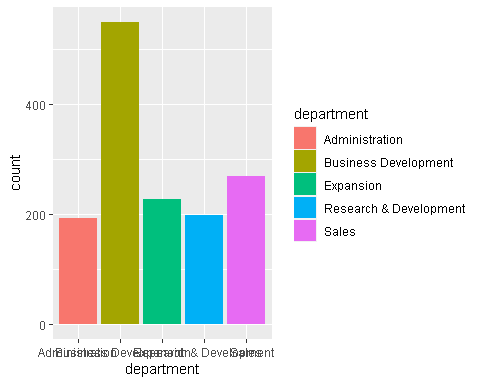
## `geom\_smooth()` using formula 'y ~ x'



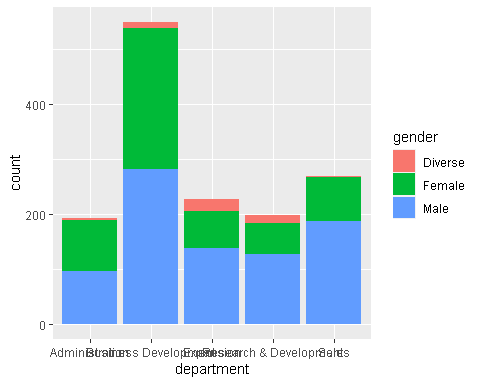
## Bar chart for gender count  
ggplot(data=salary\_data\_final) + geom\_bar(mapping=aes(x=gender, fill=gender))



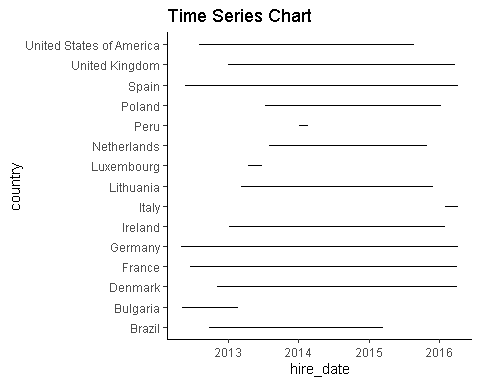
#Bar chart for department  
ggplot(data=salary\_data\_final) + geom\_bar(mapping=aes(x=department, fill=department))



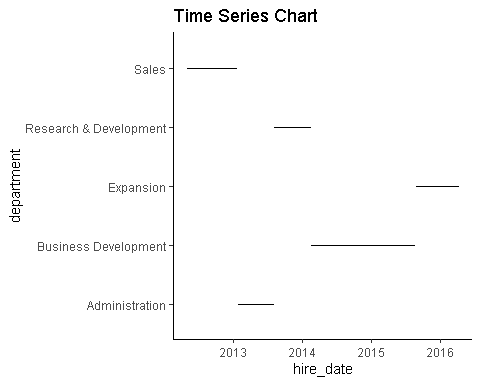
#Bar chart for department with fill as gender  
ggplot(data=salary\_data\_final) + geom\_bar(mapping=aes(x=department, fill=gender))



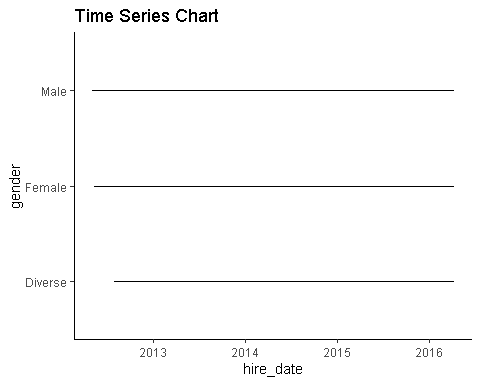
# Time Series Plot of country vs hire\_date  
theme\_set(theme\_classic())  
ggplot(salary\_data\_final, aes(x=hire\_date)) +  
 geom\_line(aes(y=country)) + labs(title="Time Series Chart")



# Time series Plot department vs hire\_date  
theme\_set(theme\_classic())  
ggplot(salary\_data\_final, aes(x=hire\_date)) +  
 geom\_line(aes(y=department)) + labs(title="Time Series Chart")



# Time series Plot gender vs hire\_date  
theme\_set(theme\_classic())  
ggplot(salary\_data\_final, aes(x=hire\_date)) +  
 geom\_line(aes(y=gender)) + labs(title="Time Series Chart")



## Findings.

Notes: There is a huge disparity in the number of diverse gender as compared to male and female gender. There is a discontinuation of hiring in countries like Peru, Luxembourg, Bulgaria and Brazil that would need to be further investigated. Huge salary difference across the different genders and countries.