DATA*6200 Assignment 1

AUTHOR
Atharva Vichare

Introduction

This document outlines steps taken for data cleaning , data manipulation and attempts to analyze the ask_a_manager data set provided .

Data Cleaning and Manipulation

Load the ask_a_manager dataset

```
library(readxl)
ask_a_manager <- read_excel("C:/Users/avich/Downloads/ask_a_manager.xlsx")
```

```
Warning: Coercing text to numeric in F26564 / R26564C6: '00'
```

Load the necessary libraries required

Renaming the columns

```
clean_ask_a_manager <- ask_a_manager |>
  rename(
    Age_group = `How old are you?`,
    Industry name = `What industry do you work in?`,
```

```
    Job_title = `Job title`,
    nothing = `If your job title needs additional context, please clarify here:`,
    Annual_salary = `What is your annual salary? (You'll indicate the currency in a later question. If you are
    Bonuses = `How much additional monetary compensation do you get, if any (for example, bonuses or overtime
    Currency = `Please indicate the currency`,
    Foreign = `If "Other," please indicate the currency here:`,
    Empty = `If your income needs additional context, please provide it here:`,
    Country = `What country do you work in?`,
    US_State = `If you're in the U.S., what state do you work in?`,
    City = `What city do you work in?`,
    Overall_Work_Exp = `How many years of professional work experience do you have overall?`,
    Work_Exp_in_field = `How many years of professional work experience do you have in your field?`,
    Education = `What is your highest level of education completed?`,
    Gender = `What is your gender?`,
    Race = `What is your race? (Choose all that apply.)`
  )
head(clean_ask_a_manager)
```

```
# A tibble: 6 × 18
  Timestamp           Age_group Industry_name    Job_title nothing Annual_salary
  <dttm>              <chr>     <chr>            <chr>     <chr>           <dbl>
1 2021-04-27 11:02:09 25-34     Education (High… Research… <NA>            55000
2 2021-04-27 11:02:21 25-34     Computing or Te… Change &… <NA>            54600
3 2021-04-27 11:02:38 25-34     Accounting, Ban… Marketin… <NA>            34000
4 2021-04-27 11:02:40 25-34     Nonprofits       Program … <NA>            62000
5 2021-04-27 11:02:41 25-34     Accounting, Ban… Accounti… <NA>            60000
6 2021-04-27 11:02:45 25-34     Education (High… Scholarl… <NA>            62000
# ℹ 12 more variables: Bonuses <chr>, Currency <chr>, Foreign <chr>,
#   Empty <chr>, Country <chr>, US_State <chr>, City <chr>,
```

```
#   Overall_Work_Exp <chr>, Work_Exp_in_field <chr>, Education <chr>,
#   Gender <chr>, Race <chr>
```

Using the rename function to rename all the columns in the data set because it is not possible to do any cleaning or analysis with the original column names because they are too long to write in the code.

Remove the duplicates

Removed the duplicates from the data set by extracting the dates and year from the Timestamp column in separate column named "Dates" and " Year" respectively because they will help analyzing how salaries vary over time and geography.

```
#Extract year and date from the Timestamp column
clean_ask_a_manager <- clean_ask_a_manager |>
  mutate(Year = format(Timestamp, "%Y"),
         Date = format(Timestamp, "%Y-%m-%d"))

#Drop the Timestamp column
clean_ask_a_manager <- clean_ask_a_manager |>
  select(-Timestamp)

#Remove Duplicates
clean_ask_a_manager <- distinct(clean_ask_a_manager)
head(clean_ask_a_manager)
```

```
# A tibble: 6 × 19
  Age_group Industry_name       Job_title nothing Annual_salary Bonuses Currency
  <chr>     <chr>               <chr>     <chr>           <dbl> <chr>   <chr>
1 25-34     Education (Higher … Research… <NA>            55000 0       USD
2 25-34     Computing or Tech   Change &… <NA>            54600 4000    GBP
3 25-34     Accounting, Bankin… Marketin… <NA>            34000 <NA>    USD
4 25-34     Nonprofits          Program … <NA>            62000 3000    USD
```

```
5 25-34     Accounting, Bankin… Accounti… <NA>            60000 7000    USD
6 25-34     Education (Higher … Scholarl… <NA>            62000 <NA>    USD
# ℹ 12 more variables: Foreign <chr>, Empty <chr>, Country <chr>,
#   US_State <chr>, City <chr>, Overall_Work_Exp <chr>,
#   Work_Exp_in_field <chr>, Education <chr>, Gender <chr>, Race <chr>,
#   Year <chr>, Date <chr>
```

Calculating total salary .

To calculate the total salary I want to add Annual_salary and Bonuses column but the Bonuses column contains NA values which cannot be added so I replaced the NA values in Bonuses column with 0 because i want to add them in Annual_salary column to find the Total_salary and having NA values in bonuses column causes problems .

```
##Replace NA Values in Bonuses with 0
clean_ask_a_manager <- clean_ask_a_manager|>
  mutate(
    Bonuses = as.numeric(Bonuses),
    Bonuses = replace_na(Bonuses, 0))
```

```
##Calculate total salaries
clean_ask_a_manager <- clean_ask_a_manager|>
  mutate(Total_salary = as.numeric(Annual_salary) + as.numeric(Bonuses))
head(clean_ask_a_manager)
```

```
# A tibble: 6 × 20
  Age_group Industry_name       Job_title nothing Annual_salary Bonuses Currency
  <chr>     <chr>               <chr>     <chr>           <dbl>   <dbl> <chr>
1 25-34     Education (Higher … Research… <NA>            55000       0 USD
2 25-34     Computing or Tech   Change &… <NA>            54600    4000 GBP
3 25-34     Accounting, Bankin… Marketin… <NA>            34000       0 USD
```

```
4 25-34     Nonprofits          Program … <NA>            62000    3000 USD
5 25-34     Accounting, Bankin… Accounti… <NA>            60000    7000 USD
6 25-34     Education (Higher … Scholarl… <NA>            62000       0 USD
# ℹ 13 more variables: Foreign <chr>, Empty <chr>, Country <chr>,
#   US_State <chr>, City <chr>, Overall_Work_Exp <chr>,
#   Work_Exp_in_field <chr>, Education <chr>, Gender <chr>, Race <chr>,
#   Year <chr>, Date <chr>, Total_salary <dbl>
```

This creates a new column named Total_salary which can be used further for analysis .

Dropping unnecessary columns

Dropped some columns from the data set which are not required for analysis and helps to enhance data clarity .

```
##Drop columns
clean_ask_a_manager <- clean_ask_a_manager |>
  select(- nothing,- Empty)
```

Cleaning the country column

I glanced through the data and found out that the country column contains irregularities which need to be standardized because the country column helps me answer salary analysis over geography as well as help to clean the currency column further.

> Limiting factor here is that in country column people have written things like global , intenration or working in asia pacific and some remote locations which makes it difficult to categorize all the entries in country names so have to use terms like Remote and International to make them stand out.

Summary of how code works :-

– Define variations using two lists [usa_variations & great_britain_variations]

– Use mutate() to modify the country column and case_when() to to set conditions and replace the string

> Please refer to the code file for code.

Currency column standardization

As mentioned previously the currency column contains currencies in various formats, abbreviations and codes which makes it difficult to analyze so I cleaned the currency column in the following steps .

Step 1 : Replacing "Other"values in the currency column

```
##Cleaning Currency column
clean_ask_a_manager <- clean_ask_a_manager|>
  mutate(
    Currency = ifelse(Currency =="Other",Foreign,Currency))
```

There are various entries with "Other" inside the currency column with its respected values in the Foreign column that can be retrived .

Step 2 : Filling missing values in Currency column by Country name

```
#Fill NA values in Currency with Country name
clean_ask_a_manager <- clean_ask_a_manager |>
  mutate (
    Currency = ifelse(is.na(Currency),Country,Currency))
```

The currency column contains various missing values which can replaced by their Country name rather than deleting the entries.

Step 3 : Standardizing currency names with str_replace_all()

The str_replae_all() is used to search for variations and replace them with the actual currency names and the regex() is used to match patterns.Dropped the 'Foreign' since now it wont be of any use.

> Plase refer to code file for the code .

Coverting currencies into one single type

After cleaning the currencies, I choose to convert all the currencies into USD($) because its not possible to do analysis with different currencies. .

Here I create a conversion_rates dataframe to store the conversion values of all the currencies .

```
#Create a conversion rates data frame to store rates
conversion_rates <- data.frame(
  Currency = c("USD", "GBP", "CAD", "EUR", "AUD/NZD", "INR", "ARS", "CHF",
               "MYR", "ZAR", "SEK", "HKD", "NOK", "BRL", "DKK", "TTD",
               "MXN", "CZK", "Bdt", "PHP", "PLN", "TRY", "CNY",
               "ILS", "AUD", "JPY", "NTD", "SGD", "KRW", "THB", "IDR",
               "NZD", "LKR", "SAR", "RM", "HRK", "NGN", "COP"),
  rate_to_usd = c(1.0, 1.36,0.74,1.47,0.64,0.012,0.0028,1.09,
                  0.24,0.052,0.092,0.13,0.086,0.19,0.15,0.15,
                  0.012,0.045,0.0095,0.018,0.23,0.036,0.14,
                  0.27,0.64,0.0067,0.032,0.74,0.0076,0.030,0.000065,
                  0.64,0.64,0.027,0.018,0.00076,0.27,0.22)
)

# Left join clean_ask_a_manager with conversion rates
clean_ask_a_manager <- clean_ask_a_manager |>
  left_join(conversion_rates, by = "Currency") |>
  mutate(
```

```
    salary_in_usd = ifelse(!is.na(rate_to_usd), Total_salary * rate_to_usd, NA)
  )

#Remove rows with salaries 0
clean_ask_a_manager <- clean_ask_a_manager |>
  filter(salary_in_usd != 0)

head(ask_a_manager)
```

```
# A tibble: 6 × 18
  Timestamp           `How old are you?` What industry do you work…¹ `Job title`
  <dttm>              <chr>              <chr>                       <chr>
1 2021-04-27 11:02:09 25-34              Education (Higher Educatio… Research a…
2 2021-04-27 11:02:21 25-34              Computing or Tech           Change & I…
3 2021-04-27 11:02:38 25-34              Accounting, Banking & Fina… Marketing …
4 2021-04-27 11:02:40 25-34              Nonprofits                  Program Ma…
5 2021-04-27 11:02:41 25-34              Accounting, Banking & Fina… Accounting…
6 2021-04-27 11:02:45 25-34              Education (Higher Educatio… Scholarly …
# ℹ abbreviated name: ¹`What industry do you work in?`
# ℹ 14 more variables:
#   `If your job title needs additional context, please clarify here:` <chr>,
#   `What is your annual salary? (You'll indicate the currency in a later question. If you are part-time or hourly,
please enter an annualized equivalent -- what you would earn if you worked the job 40 hours a week, 52 weeks a year.)`
<dbl>,
#   `How much additional monetary compensation do you get, if any (for example, bonuses or overtime in an average
year)? Please only include monetary compensation here, not the value of benefits.` <chr>,
#   `Please indicate the currency` <chr>,
#   `If "Other," please indicate the currency here:` <chr>, …
```

Here, I join the conversion_rate dataframe to my main dataframe of clean_ask_a_manager using left join and calculate the currencies in USD by simply multiplying total_salary and rates_to_usd and store the result in a new column named salary_in_usd .Further

removed the rows which contained ‘0’ in salary_in_usd column since they are practically missing plus there arent many rows so it wont affect our analysis much

Cleaning industry_name column

The industry name column has various anomalies which need to be fixed before doing analysis.

Split the strings in the industry column seperated by space, /,or and & symbol and then took the first word from every entry as the main industry .

After this used str_replace_all() which replaces repeating keywords with their original format for example industry names containing ‘IT’, Software , Computing and Tech are replaced by “IT”. The regex() is used to define regular expression for matching keywords and str_squish() removes any extra spaces ensuring the column is clean.

At last I remove the rows containing NA in the Industry_name column because they are of no use and we cannot replace them using any method since they are difficult to predict.

> Please refer to code file for the code .

> Limiting factor here is there are countless industries which belong to one group but have different meaning (ambiguity) and also there are entries where people have written all sorts of unusable things which is really difficult to pick and remove. Therefore it is hard to catrgorize all of them.

Cleaning US_State column

```
#Clean US State
clean_ask_a_manager <- clean_ask_a_manager |>
  mutate(New_US_State = case_when(
    str_detect(US State, "^California") ~ "California",
```

```
    str_detect(US_State, "^New York") ~ "New York",
    str_detect(US_State, "^Texas") ~ "Texas",
    str_detect(US_State, "^Florida") ~ "Florida",
    str_detect(US_State, "^Virginia") ~ "Virginia",
    str_detect(US_State, "^Ohio") ~ "Ohio",
    str_detect(US_State, "^Illinois") ~ "Illinois",
    str_detect(US_State, "^Georgia") ~ "Georgia",
    str_detect(US_State, "^Massachusetts") ~ "Massachusetts",
    str_detect(US_State, "^Arizona") ~ "Arizona",
    str_detect(US_State, "^Alabama") ~ "Alabama",
    str_detect(US_State, "^District of Columbia") ~ "District of Columbia",
    str_detect(US_State, "^Wyoming") ~ "Wyoming",
    str_detect(US_State, "^North Carolina") ~ "North Carolina",
    str_detect(US_State, "^Washington") ~ "Washington",
    str_detect(US_State, "^Michigan") ~ "Michigan",
    str_detect(US_State, "^Oregon") ~ "Oregon",
    str_detect(US_State, "^Nevada") ~ "Nevada",
    str_detect(US_State, "^Mississippi") ~ "Mississippi",
    str_detect(US_State, "^Colorado") ~ "Colorado",
    str_detect(US_State, "^Maryland") ~ "Maryland",
    str_detect(US_State, "^South Carolina") ~ "South Carolina",
    str_detect(US_State, "^Tennessee") ~ "Tennessee",
    str_detect(US_State, "^Pennsylvania") ~ "Pennsylvania",
    str_detect(US_State, "^Iowa") ~ "Iowa",
    str_detect(US_State, "^Arkansas") ~ "Arkansas",
    str_detect(US_State, "^Hawaii") ~ "Hawaii",
    str_detect(US_State, "^New Jersey") ~ "New Jersey",
    str_detect(US_State, "^Kentucky") ~ "Kentucky",
    str_detect(US_State, "^Indiana") ~ "Indiana",
    TRUE ~ US_State
  ))
```

Cleaned the US_state using str_detect() to answer the question about variabiltiy in salaries over geography

In this way i cleaned the whole data set and now it is ready for analysis and visualizationzation.

Data Visualization

Question 1

Which industry or industries have the highest/lowest salaries?

Industries with Highest salaries and Lowest Salaries

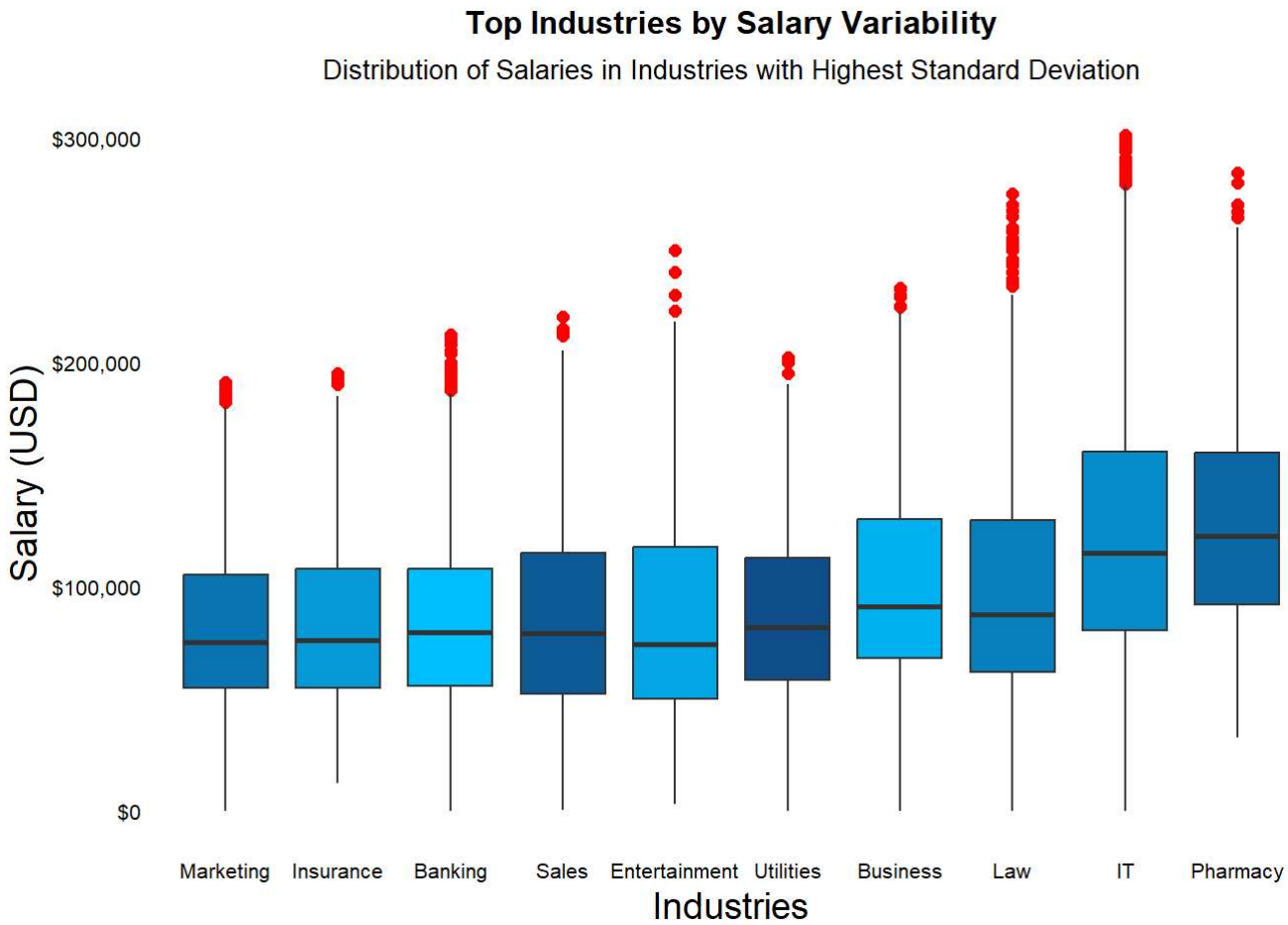
This is horizontal bar chart that ranks industries based on their average salary. Pharmacy, IT, and Law have the highest average salaries, ranging from $103k to $126k per year. Business, Engineering, and Utilities also have relatively high average salaries. Library, Social Work, and Retail have the lowest average salaries, ranging from $54k to $56k per year. Hospitality, Education, and Nonprofits also have lower average salaries. The overall salary range is quite significant, with a difference of $72k between the highest and lowest average salaries.

lowest average salaries.

Question 2 :

Which industries have the highest salary variability?

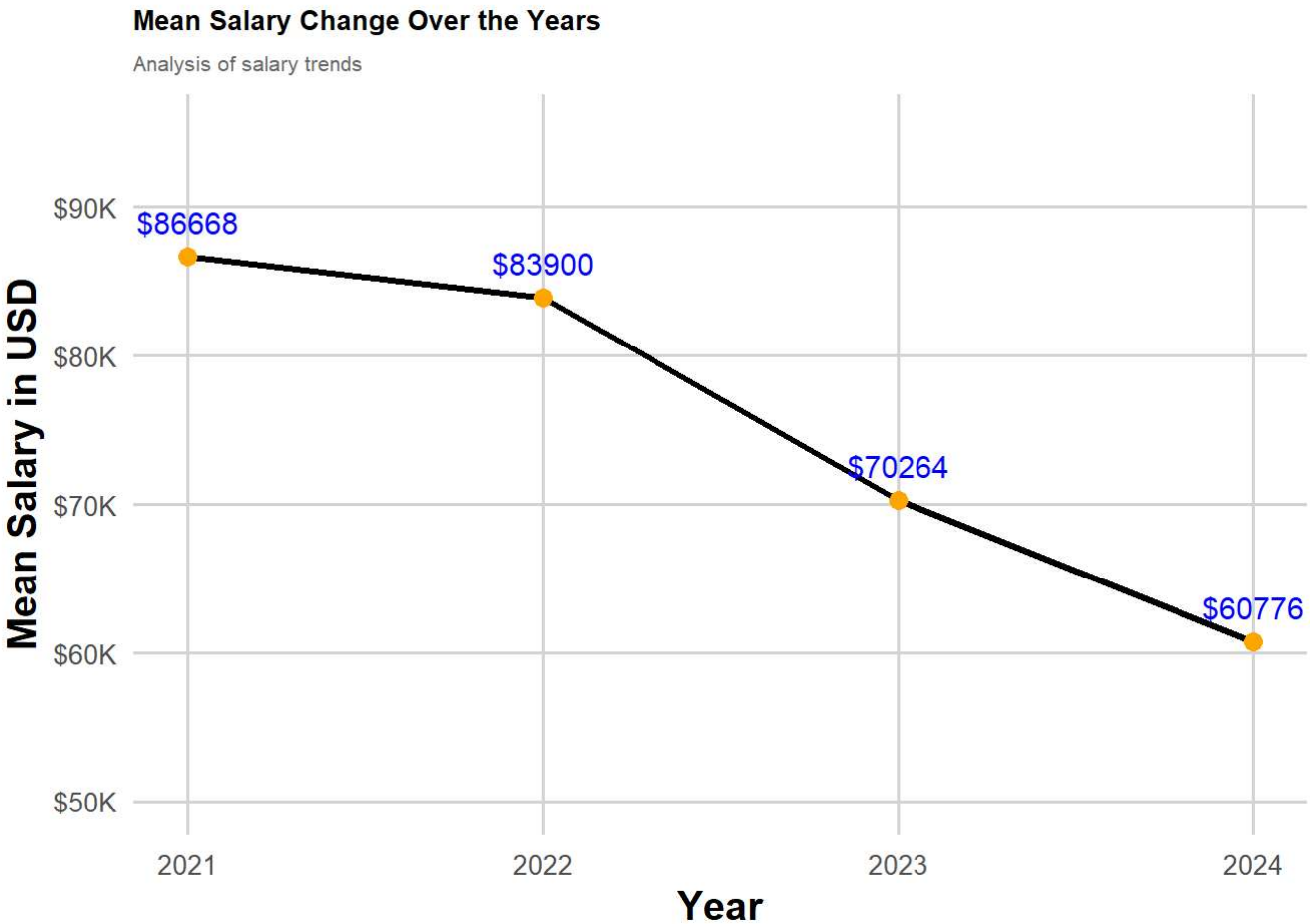
To answer this question I made use of box plot to show variability in the salaries.



The box plots for all industries show a significant spread, indicating high salary variability within each industry. Several industries have outliers, particularly Marketing and IT, which suggests that there are a small number of individuals in these industries who earn significantly more than the rest of their peers.The median salaries (the horizontal lines within the boxes) for most industries are relatively close together, suggesting that the middle 50% of earners in these industries have similar salaries. While the overall variability is high across all industries, there are some differences in the specific distributions. For example, the box plot for Pharmacy is narrower than the box plots for some other industries, indicating a smaller range of salaries within that industry. Industries like Law and Entertainment have relatively high mean salaries but also demonstrate notable variability.

Question 3:

How do salaries vary over time ?

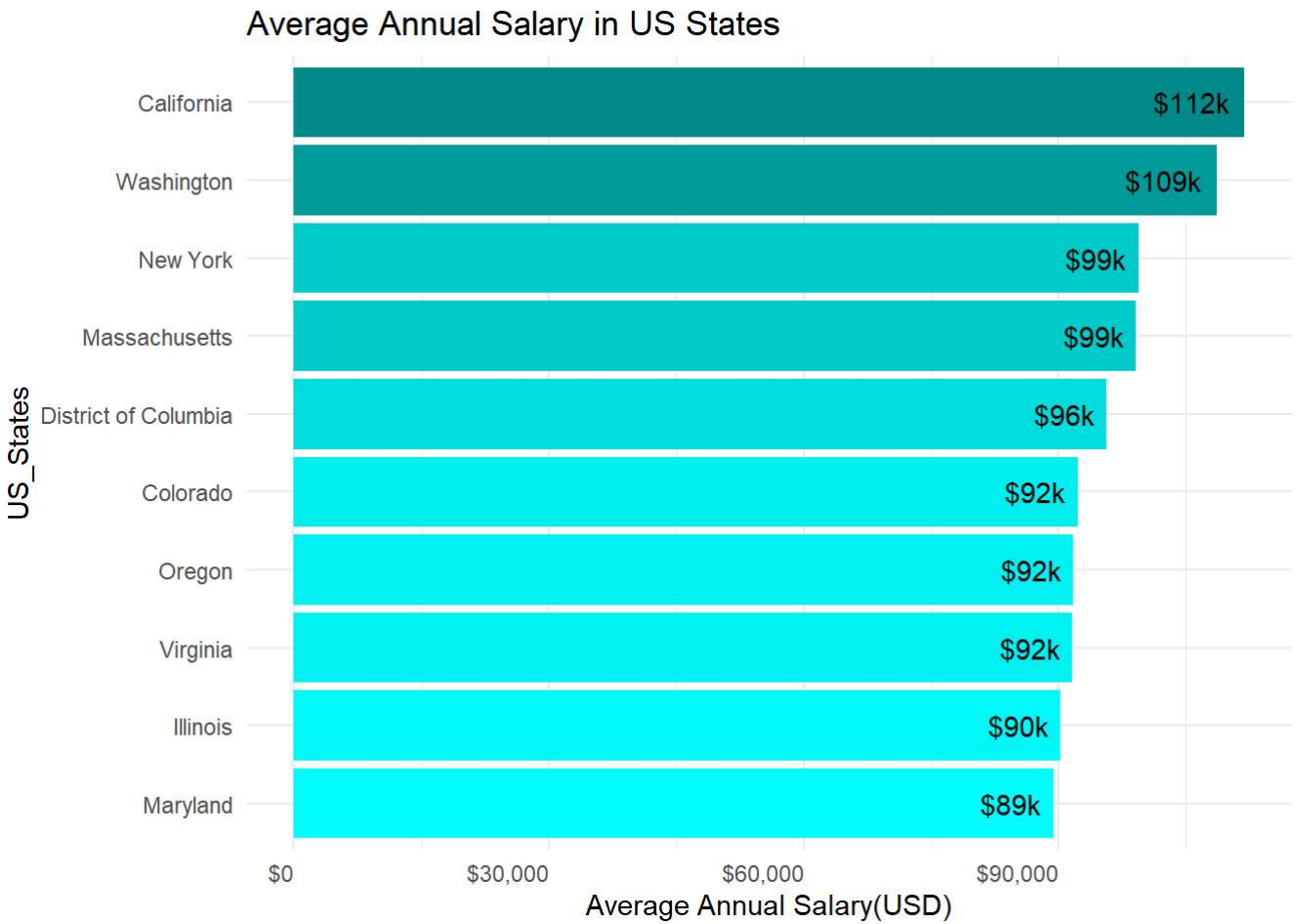


As years pass by and we come to the year 2024 we can observe that there are only 74 entries and the number of industries is much more so its not possible to say for a fact on how salaries have increased or decreased from 2023 to 2024. But still we can take a general idea of how salaries have varied.

The graph suggests that mean salary kept decreasing from 2021 to 2024 Between 2021 and 2022, the mean salary decreased from $86,631 to $83,900. Between 2022 and 2023, the mean salary decreased further from $83,900 to $70,264. Between 2023 and 2024, the mean salary decreased even further from $70,264 to $60,776. Overall, the data suggests a steady decline in mean salary over the years. This could be attributed to various factors such as economic conditions, inflation,changes in the job market due to Covid-19.

> Limit to this graph is that there should have been more data about 2024 if for certain we want to say the salaries have decreased. But since there is limited data we can only assume by looking at the graph that initially it has decreased.

How do salaries vary over geography?



This horizontal bar chart that displays the average annual salary in various US states. From the chart we can say that California has the highest average annual salary among the listed states, followed by Washington. New York, District of Columbia, Alabama, and New Jersey have similar average annual salaries, ranging from $110K to $118K. Connecticut, Virginia, and Colorado have lower average annual salaries compared to the other states.

Plot 1 : Percentage increase in salary by Overall Work Experiences

This is a line chart which shows trend in average annual salary (USD) across different work experience . The percentage values above each point shows relative percentage change in salary compared to previous experiences . Key observations :

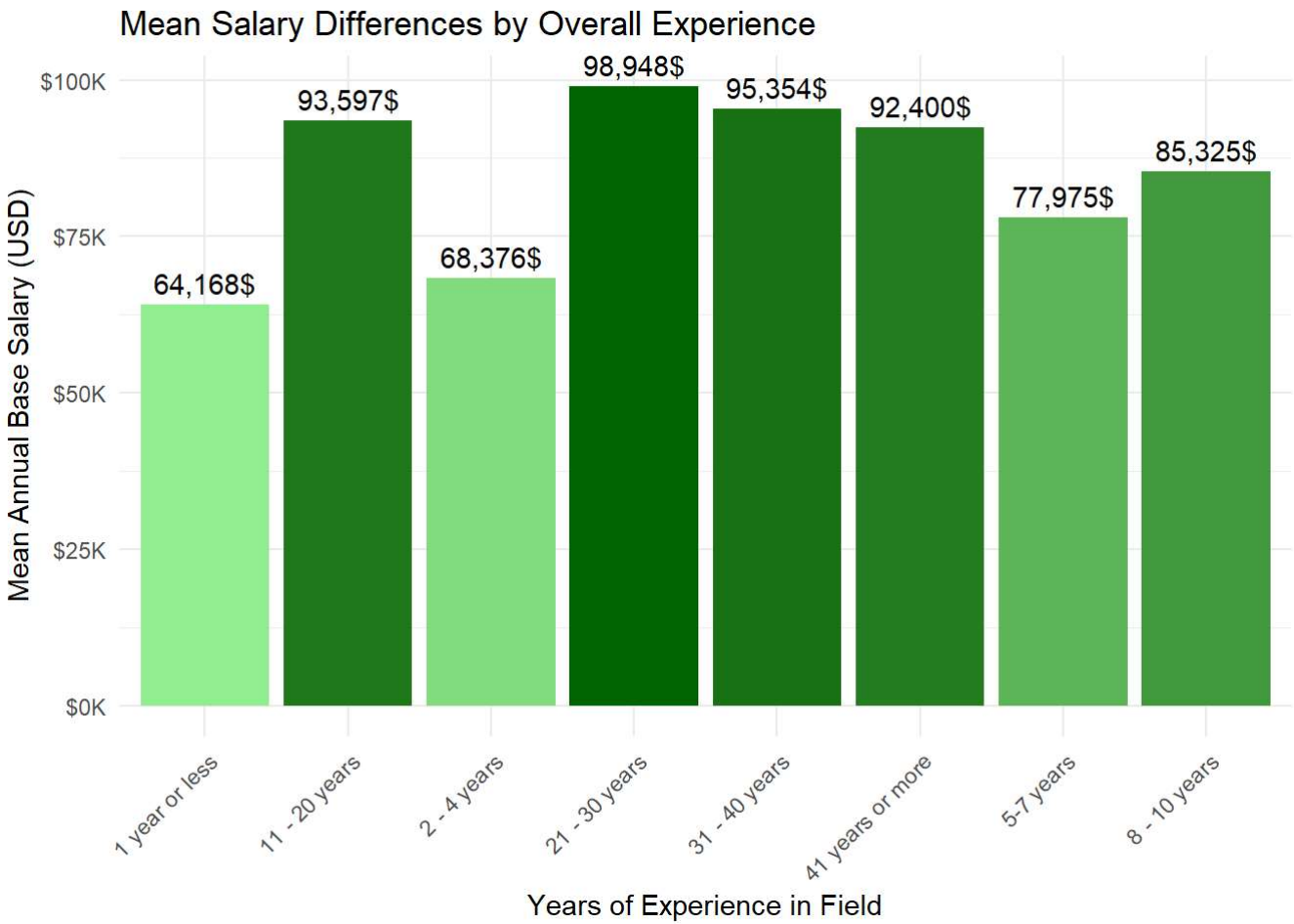
– Moving from "1 year or less" to "2 - 4 years" shows a 6.7% increase in salary

–The largest increase is from "5 - 7 years" to "8 - 10 years," at 12.5%.

– After "21 - 30 years," the salary increase rate starts decreasing, with the final category ("41 years or more") showing a slight decrease compared to "31 - 40 years."

Thus we can say that there is positive increase in salary till 21-30 years of experience and after that the trend shows decreasing trend

Plot 2 : Mean Salary differnece by Overall Experience



The graph shows the mean annual base salary differences by overall experience level. As the years of experience increase, the mean annual base salary generally increases as well.

There is a significant jump in salary between the “1 year or less” and “2-4 years” experience levels.

This suggests a significant increase in earning potential after gaining some initial experience.The salary continues to increase steadily with each additional year of experience, up until the “31-40 years” experience level.

However,after the “31-40 years” experience level, there is a slight decrease in the mean annual base salary. This could be due to factors such as retirement planning or reduced job responsibilities in later career stages.

Overall the graph suggests that higher experience leads to higher salary.