Project Summary

Problem Statement

Data science is one of the hot career topics, but how much can one expect to make a living as a Data Scientist, some articles say upward of $100K, and some suggest it can go beyond upward of $300k and more. We know there are always outliers in every market, data science is the same. Therefore, let's use some data science to find out, how much a data science realistically gets paid!

Solution

The problem focuses mainly on data exploration and interpretation, which means from a large data set, we are going to explore what matters most or least. In this project, we choose some dirty data to work with, since often time good data source is not accessible. Data cleaning is important to process in Data Science to make dirty data into functional data to be able to conduct analysis, and get the most (valuable information) with the worst is the aim of this project. After cleaning, we going to conduct hypothesis testing to let the statistic speaks for itself rather than using biased judgment.

Data Source

This data set is found on Github by Kenarapfaik, the data set was a web scrape from Selenium and BeautifulSoup. Personally, I have created projects using Selenium with chromium-browser to automate some of my personal tasks. Since it contains some of my personal login in the source code, it is not public sharable. However, this dataset is perfect for this project, it contains many useful information, but the data is quite dirty. Which required good cleaning before conducting analysis. There is a total of 3909 rows in this data set.

Programing Environment

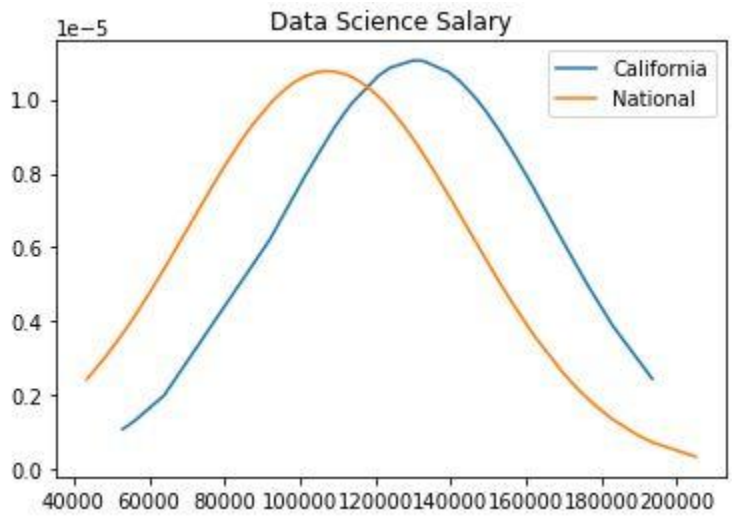
Window 10/11, Python 3.9, Jupyter Lab, Google Chrome

Libraries Used

Pandas, Statistics, Seaborn, Scipy, Matplotlib

Project Summary

Result



National Average: $107,426 /yrs

California Average: $130,799 /yrs

Since the T-Test the P-Value is 1.24e-80 which is an extremely small number of 0.00...00124 with 80sdecimal zero, which means it is much smaller than the standard 0.05 p-value, we will reject the Ho: Null Hypothesis and accept the Ha: Alternative Hypothesis, such that it is concluded that California is significantly different than the overall US average in data science salary.

Project Link and Source Code

Main Analysis Work in IPYNB:

https://github.com/GiggleSamurai/Data-Cleaning-Hypothesis-Analysis-on-Dirty-Data/blob/a8b17efafa4cc35eededdba7937c515f68128bba/Salary%20Analyst%20Project.ipynb