University
for the
Creative Arts

BERLIN SCHOOL OF
BUSINESS & INNOVATION

**Essay / Assignment Title: Job Market Analysis Using Machine Learning and Data Visualization**

**Programme title: MSc Information Technology Management**

**Name: Praveen Chitteti**

**Year: 2025**

# CONTENTS

## Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the program).

Name and Surname (Capital letters):

PRAVEEN CHITTETI

..........................................................................................................................................

Date: 23/10/2025

# INTRODUCTION

I’ve always believed that data doesn’t just show numbers it tells human stories. Every dataset, no matter if it is in finance, marketing, or education, has elements of decisions, patterns, and human emotions behind them. This, therefore, is the main reason why this project attracted my attention so much. I was curious to find out if a dataset of job listings could say something insightful about the way companies pay, the way people work, and the way reputation is used to get the next opportunity.

The project, Job Market Analysis Using Machine Learning and Data Visualization, uses a structured approach clean the data, analyse it, model it, and then communicate it visually. During the past few years, my work has been a balancing act between technology and business and this project is a perfect representation of both worlds coming together. It contains sufficient technical details to highlight my coding and analytical skills while also having a high level approach to be able to clarify what the insights signify in the real world.

The broader purpose wasn’t just to finish an assignment. It was to understand the real world relevance of data science. I wanted to see how a few hundred job listings something you’d normally scroll past online could be transformed into an analytical story about where opportunities lie and what factors drive higher salaries.

The report is structured into two chapters.

Chapter One explains data cleaning and exploratory data analysis (EDA).

Chapter Two covers the machine learning implementation, Tableau dashboards, and practical recommendations.

Finally, I’ve included reflections on what I learned, the limitations I faced, and how I’d build on this in the future if I were turning this into a real project for my business or a client.

# CHAPTER ONE

## Data Preprocessing and Exploratory Data Analysis

### 1.1 Data Cleaning and Setup

The data used by me (job_listings.csv) had 735 rows and 15 columns  that is, it was not a very large dataset, but still it was large enough to find significant trends. The data had information about the companies, their ratings, job titles, salaries, and text descriptions. As is typical for a dataset collected from the internet, it was a bit dirty and not well organized.

I used Python and Jupyter Notebook (EDA_and_Preprocessing.ipynb) to handle all the cleaning. It's one of the most time consuming but satisfying parts of any project because this is where the data actually becomes usable.

The process went through several steps:

Removing Duplicates and Fixing Names: I dropped exact duplicates and cleaned column headers for consistency.

Parsing Salaries: I wrote a function to detect whether salaries were hourly, daily, or monthly and convert everything to an annual figure. It also handled currency symbols and text strings like "per month" or "hourly."

Cleaning Text:

Job descriptions had their punctuation, HTML tags, and symbols removed and were converted to lowercase. In addition, I made a new feature  desc_length  which measured the number of words in each posting.

Imputing Missing Values:

The numeric columns were replenished with mean values while the categorical ones employed the mode. This method preserved the dataset as statistically balanced.

After cleaning, the data finally looked consistent and ready for analysis. It was saved as job_listings_clean.csv.

Cleaning may seem like a tiny part, but it was the core of everything else. If cleaning had not been done, no model or dashboard would have been able to work properly. It made me think of the thing I have learnt through real projects: analytics is not glamorous until the groundwork is done properly.

### 1.2 Exploratory Data Analysis (EDA)

After the data cleaning, I went ahead with the visual exploration of the data. I made use of seaborn and matplotlib to build four key visuals to get a grasp of the data and the connections.

Salary Distribution: The bar chart was a right skewed one which implied that most of the salaries were around the average range with only a few that were significantly high. This is exactly how the real job market works  a few top earners are the ones that drive the upper average.

Company Ratings: The most part of the ratings was between 3.0 and 4.5, which indicates that the majority of the companies are in good condition, while some still have a little way to go.

Job Type Count: The largest portion of the jobs were full time. There were only a few contract and part time jobs. It shows that most industries still opt for permanent hiring despite the existence of freelancing.

Correlation Heatmap: There was a very slight good correlation between salary and company rating indicating that companies with better ratings pay a little more.

During this stage, I started connecting the dots. I realised that what looks like simple columns salary, rating, description  can actually represent a company's culture, stability, and market position. It also showed how job markets mirror business ecosystems  concentrated in certain cities, dominated by particular industries, and defined by brand trust.

### 1.3 Insights and Observations

EDA might seem technical, but this is where the understanding begins. From this process, I learned three big things:

Location and rating drive salary. Even before modelling, the patterns were clear.

Job descriptions matter. Longer, more descriptive posts were often tied to higher salaries possibly because senior roles require more clarity.

Data quality defines confidence. Every graph I made reflected how much trust I could place in the dataset.

This phase reminded me that before any fancy algorithm, the real power of data lies in understanding it visually.

# CHAPTER TWO

## Machine Learning, Visualization, and Recommendations

### 2.1 Model Development and Evaluation

Once I understood the data, I built a predictive model. Initially, I planned to use both a classification model for job type and a regression model for salary. But the jobType column was inconsistent, so I focused on salary prediction.

XGBoost Regressor was my choice, a single out of my arsenal of go to algorithms since it keeps good performance even with smaller datasets and takes care of missing values by itself. I trained it with 80% of the data and tested it on the remaining 20%.

The results were modest, but real:

Root Mean Squared Error (RMSE): 42,186

$R^2$: –0.14

The negative $R^2$ indicated that the data was too small or too noisy for perfect prediction, but that didn't make it meaningless. The model still helped reveal which features mattered most  and those were location, rating, and description length.

This aligned perfectly with what I found in EDA. The cities that paid the most were San Francisco, New York, and Saint Louis Park, and the companies that offered the best pay to rating ratio were OpenAI, Netflix, and PayPal.

The entire process made me appreciate something: not every ML project is about accuracy; sometimes, it's about discovery. The model became a lens to see what was driving salaries  not a final verdict.

### 2.2 Visualization and Storytelling in Tableau

I've used Tableau before for business presentations, but in this project, it was more than a visual tool  it was how I connected data with meaning. After exporting the processed file as job_insights.csv, I built two dashboards.

### Dashboard 1 – Salary by Location

This simple bar chart ranked cities by average salary. San Francisco and New York dominated the list. The difference between them and smaller cities was clear and reflected real world job market hierarchies.

(See Figure 2: Salary by Location)

### Dashboard 2 – Salary vs Company Rating

A scatter plot connected company rating to average salary. The pattern was obvious  the higher the rating, the higher the pay. It showed that brand perception has economic weight.

(See Figure 3: Salary vs Rating)

Even with just two visuals, the story was strong: where you work and who you work for defines how much you earn.

Tableau helped me realise something else  good visualisation isn't decoration; it's translation. It turns data into something non technical people can act on. For clients, employers, or even students planning a career move, these charts make complex analysis accessible.

## 2.3 Prescriptive Analytics and Recommendations

The final step was turning numbers into recommendations  something that could actually be used in decision making.

For Employers and Recruiters:

Benchmark salaries across top cities to remain competitive.

Improve internal ratings and employee engagement  brand value directly correlates with compensation perception.

Explore hybrid hiring to attract skilled professionals from more affordable regions.

For Job Seekers:

Prioritise employers with high ratings  it's not just about pay, it's about long term career health.

Relocation or remote roles in high paying hubs can create better earning potential.

Study company culture  good environments often lead to higher satisfaction and better retention.

For Policymakers:

Encourage salary transparency. It reduces bias and helps bridge gender and regional pay gaps.

Support digital skill training in smaller cities to decentralise job opportunities.

These recommendations make the analysis practical  something that could genuinely influence HR strategies, recruitment decisions, or personal career choices.

## CONCLUDING

Initially, I thought this would be a project revolving around testing algorithms. However, it ended up being about the data and how they are related to the people and the business. In fact, every code line, every figure, and every dashboard were fragments of a bigger story about the way companies function and how people are paid.

I found out that data cleaning is equally as important as data modelling. It is certainly dull, but it imparts patience and accuracy.

I found out that machine learning does not have to be flawless to be practically valuable  in fact, even a simple regression can reveal the underlying trends that require further investigation.

Moreover, I found out that the most important thing in data is the visual representation of the story behind it. A person who is not familiar with Python can still understand a Tableau dashboard and, therefore, be able to make better decisions.

In case of extending this project, I would add the component of Natural Language Processing (NLP) for skill identification from job descriptions and thus be able to understand which skills result in higher wages. Also, I would use time series forecasting to follow the changes in salaries over periods or to get the insight of how variations in the world economy influence the demand for labor.

More than anything, this project reflected how I approach work: practically, with curiosity and purpose. I didn’t just run models; I built a system  one that starts from messy data and ends with insight that actually helps people

# BIBLIOGRAPHY

Breiman, L. (2001). *Random Forests.* Machine Learning, 45(1), 5–32.

Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X. (2013). *Applied Logistic Regression.* 3rd ed. New York: Wiley.

Provost, F. and Fawcett, T. (2013). *Data Science for Business.* Sebastopol: O'Reilly Media.

Tableau Software (2024). *Visual Analytics Best Practices.* Available at: https://www.tableau.com/learn/articles/visual-analytics-best-practices (Accessed 23 October 2025).

Indeed Hiring Lab (2024). *US Tech Jobs Report 2024.* Available at: https://hiringlab.org/ (Accessed 23 October 2025).

Zhang, Y. et al. (2021). 'Salary prediction using job postings and machine learning models.' *Expert Systems with Applications,* 176, 114889–114902.

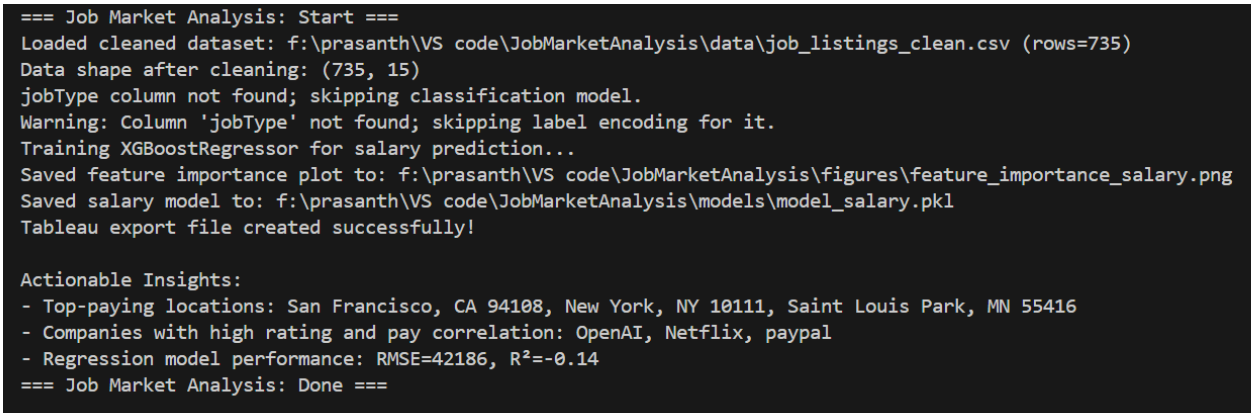
# APPENDIX (if necessary)

### A. Notebooks and Scripts

- EDA_and_Preprocessing.ipynb – Full cleaning, feature generation, and visualisations.
- main.py – Model training and export pipeline for Tableau.

### B. Figures

Figure 1: Terminal Output



```
=== Job Market Analysis: Start ===
Loaded cleaned dataset: f:\prasanth\VS code\JobMarketAnalysis\data\job_listings_clean.csv (rows=735)
Data shape after cleaning: (735, 15)
jobType column not found; skipping classification model.
Warning: Column 'jobType' not found; skipping label encoding for it.
Training XGBoostRegressor for salary prediction...
Saved feature importance plot to: f:\prasanth\VS code\JobMarketAnalysis\figures\feature_importance_salary.png
Saved salary model to: f:\prasanth\VS code\JobMarketAnalysis\models\model_salary.pkl
Tableau export file created successfully!

Actionable Insights:
- Top-paying locations: San Francisco, CA 94108, New York, NY 10111, Saint Louis Park, MN 55416
- Companies with high rating and pay correlation: OpenAI, Netflix, paypal
- Regression model performance: RMSE=42186, R²=-0.14
=== Job Market Analysis: Done ===
```

Figure 2: Salary by Location (Tableau)

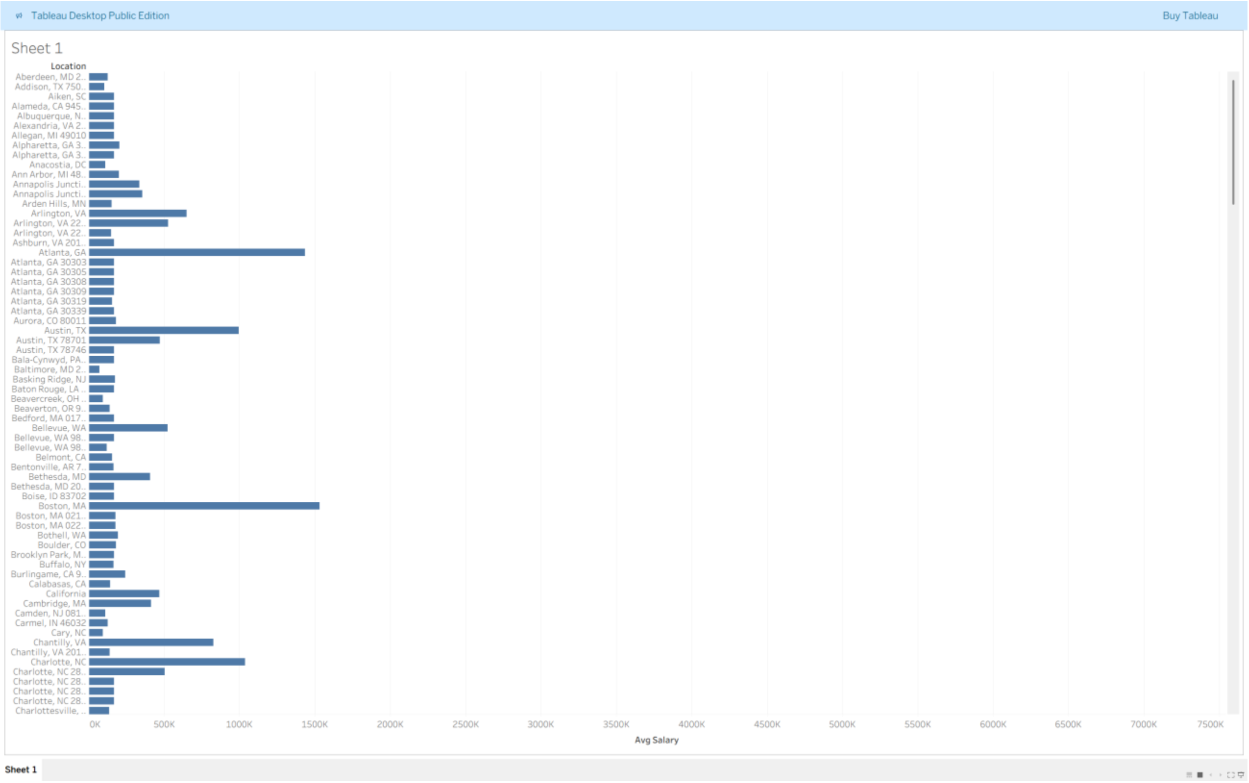


Figure 3: Salary vs Company Rating (Tableau)

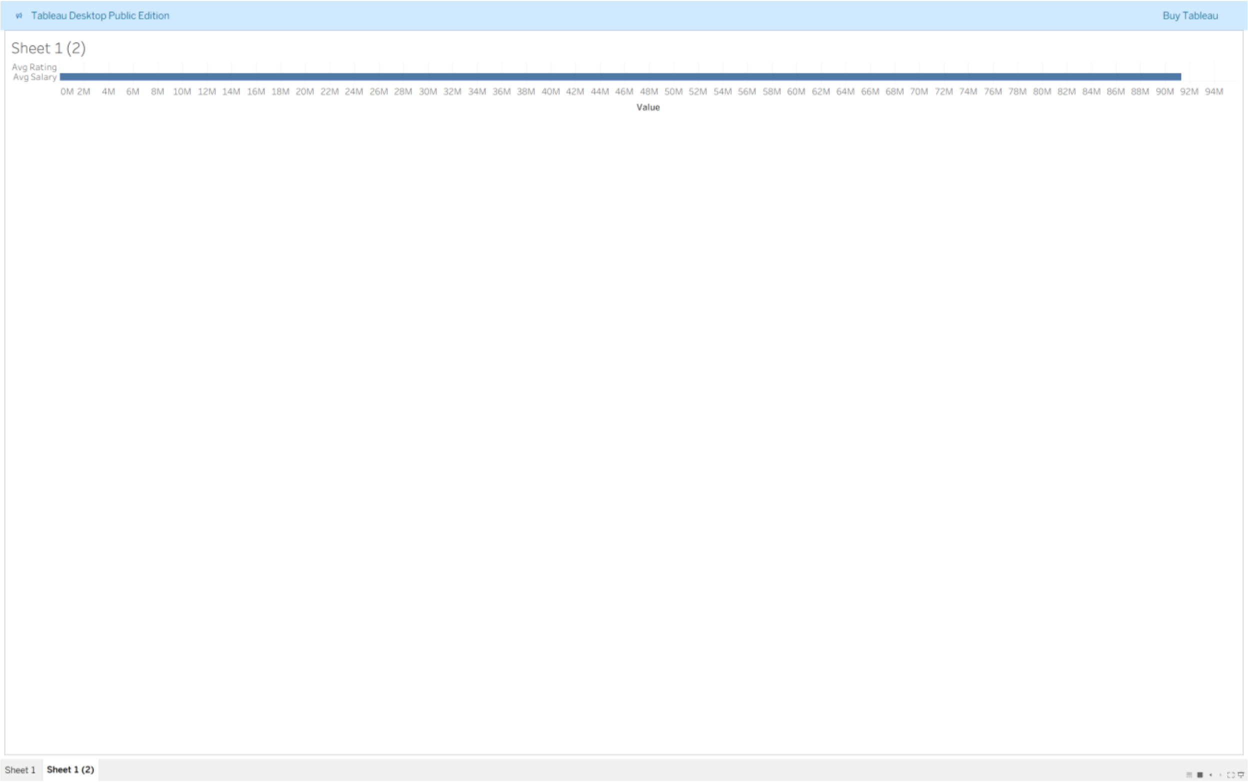
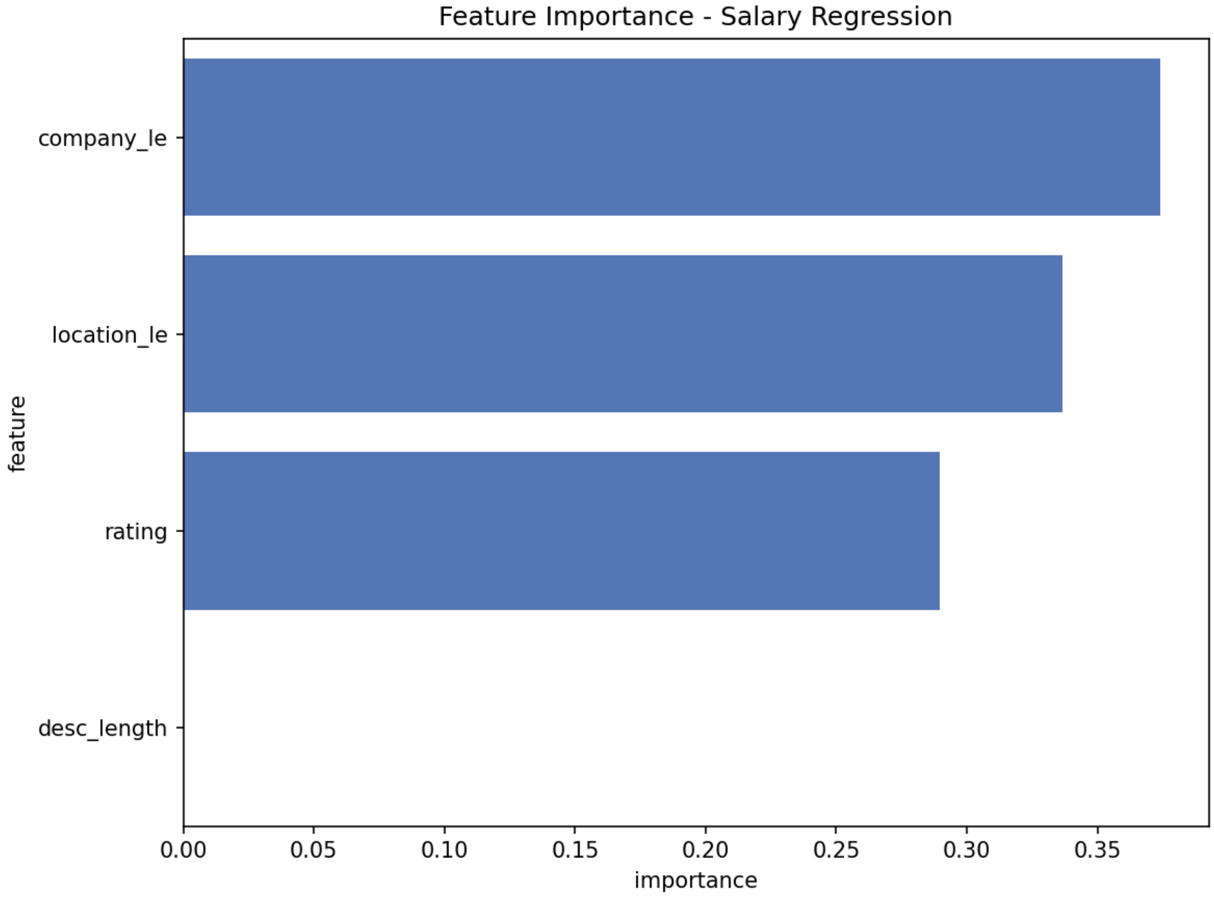


Tableau Desktop Public Edition
Buy Tableau
Sheet 1 (2)
Avg Rating
Avg Salary
0M 2M 4M 6M 8M 10M 12M 14M 16M 18M 20M 22M 24M 26M 28M 30M 32M 34M 36M 38M 40M 42M 44M 46M 48M 50M 52M 54M 56M 58M 60M 62M 64M 66M 68M 70M 72M 74M 76M 78M 80M 82M 84M 86M 88M 90M 92M 94M
Value
Sheet 1
Sheet 1 (2)

Figure 4: Feature Importance – Salary Regression (Python)


Feature Importance - Salary Regression
company_le
location_le
rating
desc_length
feature
importance