EMPLOYEE PERFORMANCE AND STATISFACTION ANALYSIS

By
MARIANIHILL A

Data Science Domain

TABLE OF CONTENT

LIST OF FIGURES

ABSTRACT

This project focuses on developing an interactive dashboard to analyse employee performance and satisfaction using a detailed dataset. The dataset includes key variables such as employee demographics, performance metrics, feedback scores, and salary information. The primary goal is to provide HR professionals and managers with actionable insights to enhance employee engagement and organizational efficiency.

The analysis begins with data cleaning and exploration, including checking for missing values and duplicates. Key aspects such as age and gender distributions are examined to understand their impact on productivity and satisfaction. Performance metrics are analysed to reveal correlations between project completion rates, productivity, and satisfaction levels. Basic machine learning techniques, including regression models, are applied to explore relationships between salary, feedback scores, and performance outcomes.

Interactive visualizations, including histograms, scatter plots, and bar charts, are used to present findings clearly. The dashboard highlights significant trends and patterns, such as the effect of salary on productivity and the relationship between feedback scores and performance. This project offers a comprehensive tool for visualizing and interpreting data, ultimately supporting data-driven decisions to improve employee satisfaction and enhance overall organizational performance.

CHAPTER-1
INTRODUCTION

In today’s competitive business environment, understanding the factors that drive employee performance and satisfaction is essential for organizational success. This project focuses on creating an interactive dashboard to analyse and visualize data related to employee performance and satisfaction. By leveraging a comprehensive dataset that includes employee demographics, performance metrics, feedback scores, and salary information, the project aims to uncover valuable insights that can help HR professionals and managers make informed decisions.

The dashboard will provide a detailed exploration of key variables such as age, gender, project completion rates, and satisfaction levels. Through data visualization and basic machine learning techniques, the project will identify patterns and correlations between these factors. For instance, how different salary ranges and feedback scores impact productivity and satisfaction will be examined. Additionally, the project will explore departmental and positional variations in performance metrics.

This approach enables a clearer understanding of how various elements influence employee engagement and efficiency. By providing a user-friendly interface for data exploration, the dashboard will assist in identifying areas for improvement and implementing strategies to enhance overall employee well-being and organizational performance.

CHAPTER-2
PROBLEM STATEMENT

Organizations face ongoing challenges in understanding the various factors that influence employee performance and satisfaction. HR professionals and managers gather large amounts of data, including employee demographics, performance metrics, feedback scores, and salary information. While this data is valuable, the sheer volume and complexity can make it difficult to analyse effectively, limiting the ability to identify key trends and insights that could lead to improved employee engagement, satisfaction, and overall productivity.

The traditional methods of analysing employee performance and satisfaction data are often manual, time-consuming, and prone to errors, making it hard to draw meaningful conclusions. Without proper tools to visualize and explore the data interactively, HR teams and managers may miss important correlations and patterns. For example, it may be difficult to determine how salary impacts employee satisfaction or how the number of projects completed correlates with productivity. Similarly, feedback provided by employees may not be analysed in a way that highlights key areas for improvement.

This project seeks to address these issues by developing an interactive dashboard that leverages the provided dataset to explore and analyse factors impacting employee performance and satisfaction. The dashboard will provide a clear visual representation of the data, enabling users to view insights on age and gender distributions, correlation among performance metrics, feedback scores, salary versus productivity, and much more. By offering an intuitive interface for HR professionals and managers to explore these factors, the project will help them make informed decisions that enhance employee engagement and organizational efficiency.

Ultimately, the goal is to equip HR teams with a tool that streamlines the data analysis process, allowing them to act quickly on insights that improve employee well-being, satisfaction, and product company.

CHAPTER – 3

TECHNOLOGY ADOPTED

3.1. Machine Learning

Machine learning is a specialized area within the broader field of artificial intelligence (AI) that focuses on enabling machines to mimic intelligent human behavior. It achieves this by leveraging historical data to train algorithms, which then use this knowledge to make predictions or classify new data. Through a process of continuous learning and adaptation, machine learning algorithms refine their models and improve their ability to make accurate predictions over time. This iterative learning approach distinguishes machine learning from traditional programming, as the algorithms themselves adjust their internal parameters based on the data they receive. By harnessing the power of data, machine learning facilitates the automation of tasks and enables systems to make informed decisions without explicit programming.

CHAPTER – 4

DETAILS OF TOOLS USE

4.1 Python

Python is a popular computer programming language used to create software and websites, automate processes, and analyze data. Python is a universal language, which meansit may be used to make many various types of applications and isn't tailored for any particular issues. Its adaptability and beginner-friendliness have elevated it to the top of the list of programming languages in use today.

Python is frequently used for creating websites and applications, automating repetitivetasks, and analyzing and displaying data. Python has been used by non-programmers, including accountants and scientists, for a range of routine activities including managing finances since it is very simple to learn.

4.2 Data Analysis and Machine Learning in Python

Python provides a rich set of libraries and tools for data analysis and traditional machinelearning. NumPy and Pandas are fundamental libraries that enable efficient numerical operations, data manipulation, and statistical computations. These libraries simplify tasks suchas data transformation, filtering, grouping, and descriptive statistics, empowering data analyststo explore and analyze data effectively.

For visualization, Matplotlib and Seaborn are powerful libraries in Python. Matplotlib offers awide range of plotting capabilities, while Seaborn provides high-level statistical visualizations.These libraries enable data analysts to create visually appealing and informative graphs, charts, and plots to communicate insights effectively.

Scikit-learn is a comprehensive machine-learning library in Python. It provides a wide range of algorithms for classification, regression, clustering, and dimensionality reduction. Scikit- learn also offers tools for model evaluation, hyperparameter tuning, and feature selection. Its user-friendly API and extensive documentation make it accessible to both beginners and experienced practitioners.

4.3 Libraries Used

4.3.1.Pandas

Pandas was used for data manipulation and analysis, particularly for loading, cleaning, and organizing the dataset. It allowed for easy management of data structures such as DataFrames and provided functionality to handle missing data and perform exploratory data analysis.

4.3.2.NumPy

NumPy was employed for numerical computing, offering support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions. It facilitated efficient computation, particularly for operations involving the manipulation of arrays.

4.3.3.Scikit-learn

Scikit-learn was the primary machine learning library used in the project. It provided tools for splitting the dataset, training the classification model (Random Forest and Logistic Regression), and evaluating the model's performance through metrics such as accuracy, confusion matrix, and classification reports.

4.3.4.Matplotlib

Matplotlib was used to create static visualizations that represented the data and model performance. This included bar charts, histograms, and scatter plots to visually examine feature distributions and relationships.

4.3.5.Seaborn

Seaborn, built on top of Matplotlib, was used for more complex and aesthetically enhanced visualizations. It provided tools for generating heatmaps, pair plots, and correlation matrices, which helped in understanding feature relationships and visualizing model performance.

CHAPTER-5
SOLUTION FOR THE PROBLEM

To address the challenge of analysing employee performance and satisfaction, the provided dataset will be explored using both basic machine learning and visualization techniques. First, the data will be cleaned by checking for missing values, duplicates, and assessing the overall structure. The analysis will start with understanding employee demographics, such as age and gender distributions, and their impact on productivity and satisfaction rates.

Next, performance metrics such as the number of projects completed, productivity, and satisfaction rates will be analysed. Correlation analysis will be performed to identify relationships between these metrics. A simple regression model will be used to evaluate how different factors, such as salary and feedback scores, influence productivity and satisfaction.

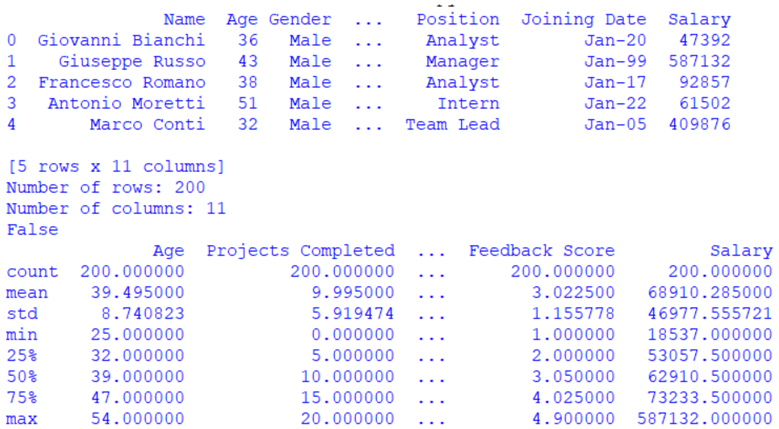
Visualizations like histograms, scatter plots, and bar charts will help in identifying trends and patterns. For example, age, department, and position distributions will be examined to highlight how different roles and salary ranges affect performance. Basic Natural Language Processing (NLP) will be used to analyse feedback data, categorizing it into positive, negative, or neutral sentiments, providing actionable insights into employee sentiments.

The goal of this solution is to provide HR professionals with a comprehensive, interactive dashboard that visualizes key insights and helps in decision-making to enhance employee engagement and organizational performance.

5.1 Data Overview

The data overview process involves first reading the dataset to understand its structure. Next, checking for duplicates ensures data integrity, followed by identifying missing values that may impact the analysis. Finally, data distributions are examined to visualize how variables like age, salary, and satisfaction are spread across the dataset.

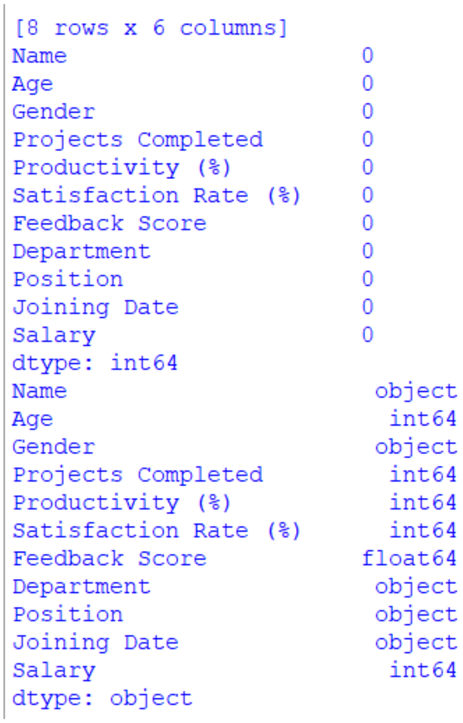
1. Read data
2. Check for Duplicates
3. Check for Missing Values
4. Data Distributions



```
               Name  Age Gender  ...   Position  Joining Date  Salary
0  Giovanni Bianchi   36   Male  ...    Analyst        Jan-20   47392
1    Giuseppe Russo   43   Male  ...    Manager        Jan-99  587132
2  Francesco Romano   38   Male  ...    Analyst        Jan-17   92857
3   Antonio Moretti   51   Male  ...     Intern        Jan-22   61502
4       Marco Conti   32   Male  ...  Team Lead        Jan-05  409876

[5 rows x 11 columns]
Number of rows: 200
Number of columns: 11
False
              Age  Projects Completed  ...  Feedback Score         Salary
count  200.000000          200.000000  ...      200.000000     200.000000
mean    39.495000            9.995000  ...        3.022500   68910.285000
std      8.740823            5.919474  ...        1.155778   46977.555721
min     25.000000            0.000000  ...        1.000000   18537.000000
25%     32.000000            5.000000  ...        2.000000   53057.500000
50%     39.000000           10.000000  ...        3.050000   62910.500000
75%     47.000000           15.000000  ...        4.025000   73233.500000
max     54.000000           20.000000  ...        4.900000  587132.000000
```

Fig 5.1.1 Read Data and Check for Duplicates



```
[8 rows x 6 columns]
Name                        0
Age                         0
Gender                      0
Projects Completed          0
Productivity (%)            0
Satisfaction Rate (%)       0
Feedback Score              0
Department                  0
Position                    0
Joining Date                0
Salary                      0
dtype: int64
Name                      object
Age                        int64
Gender                    object
Projects Completed         int64
Productivity (%)           int64
Satisfaction Rate (%)      int64
Feedback Score           float64
Department                object
Position                  object
Joining Date              object
Salary                     int64
dtype: object
```

Fig 5.1.2 Check for Missing Values

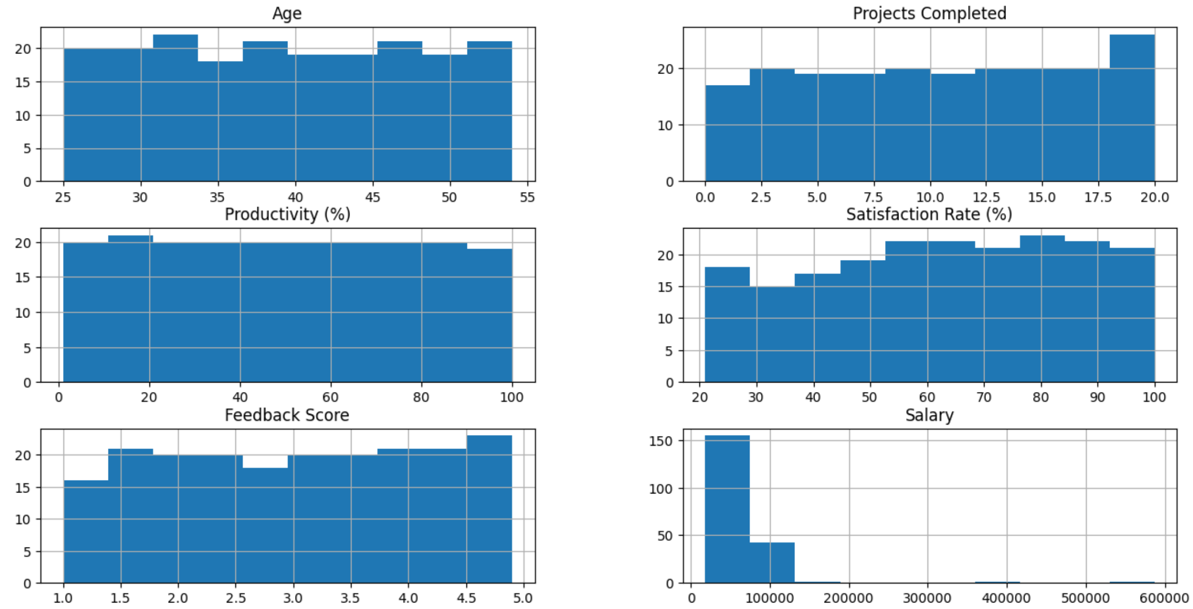


Fig 5.1.3 Data Distributions

5.2 Employee Distribution

Employee distribution focuses on analyzing age and gender across the dataset. Age distribution visualizes the range and frequency of different age groups, helping to identify trends related to employee demographics. Gender distribution examines the proportion of male and female employees, providing insights into the organization's workforce diversity.

1. Age Distribution
2. Gender Distribution

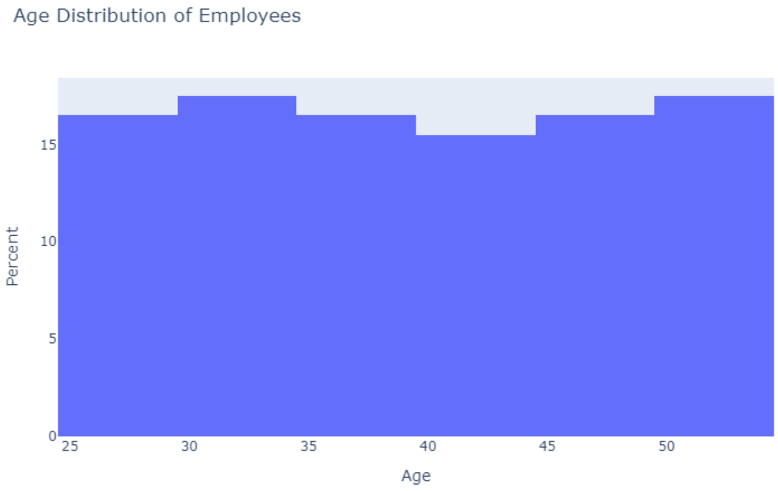


Fig 5.2.1 Age Distribution

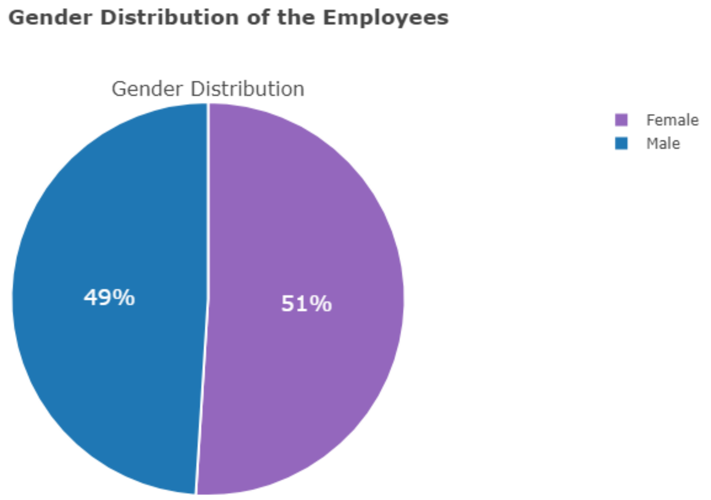


Fig 5.2.2 Gender Distribution

5.3 Department and Position Analysis

The Department and Position Analysis examines how employees are distributed across various departments and positions within the organization. It further analyzes the average productivity and satisfaction rates for each department and position, providing insights into which areas may excel or need improvement in employee performance and well-being.

1. Employee Distribution by Department and Position
2. Average Productivity and Satisfaction rate by Department and Position

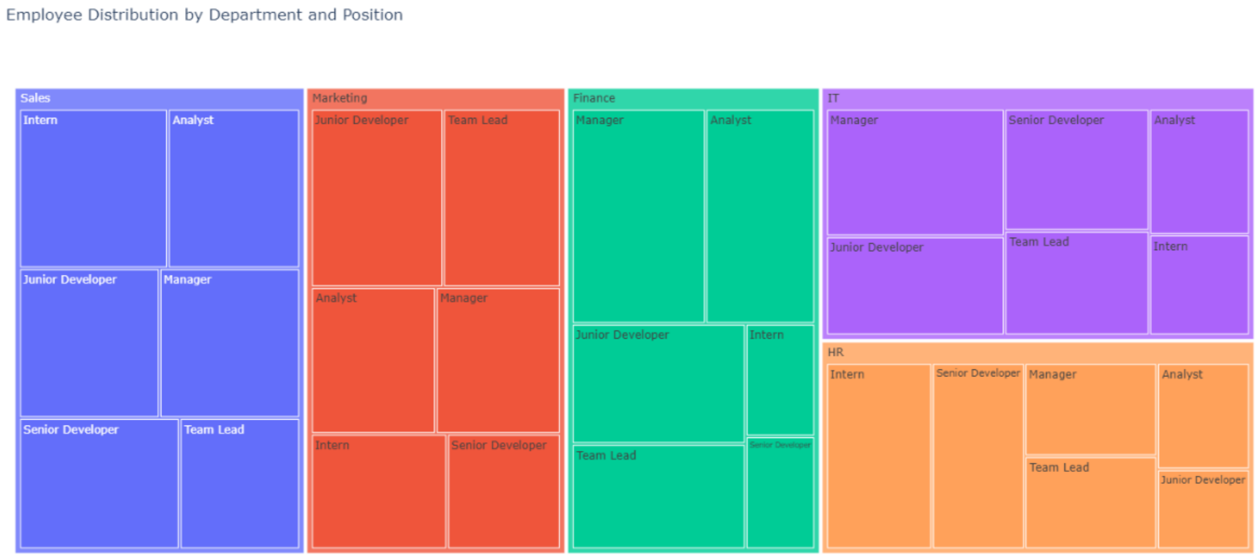


Fig 5.3.1 Employee Distribution by Department and Position

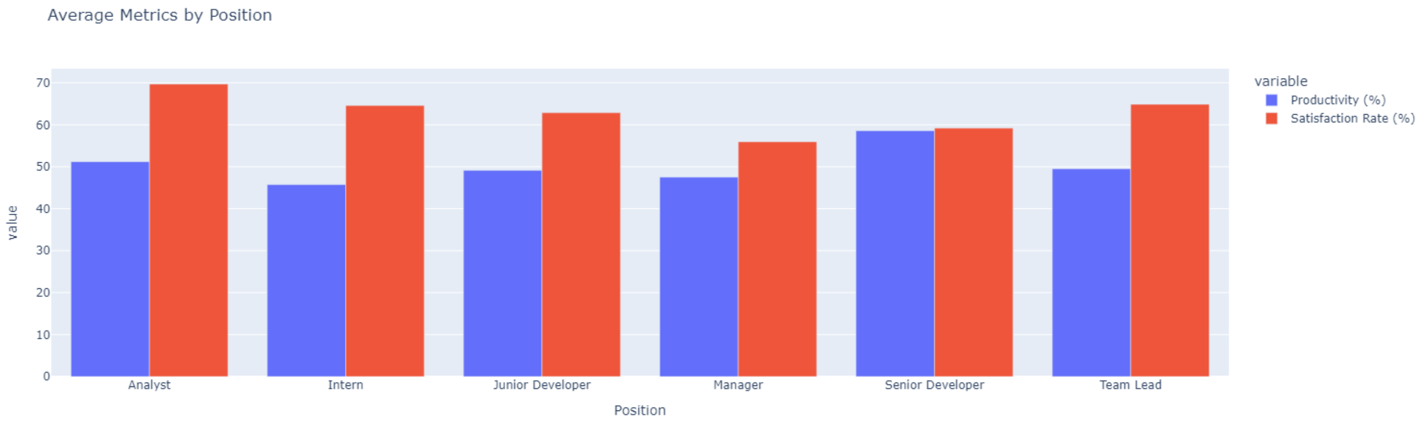


Fig 5.3.2 Average Productivity and Satisfaction rate by Position

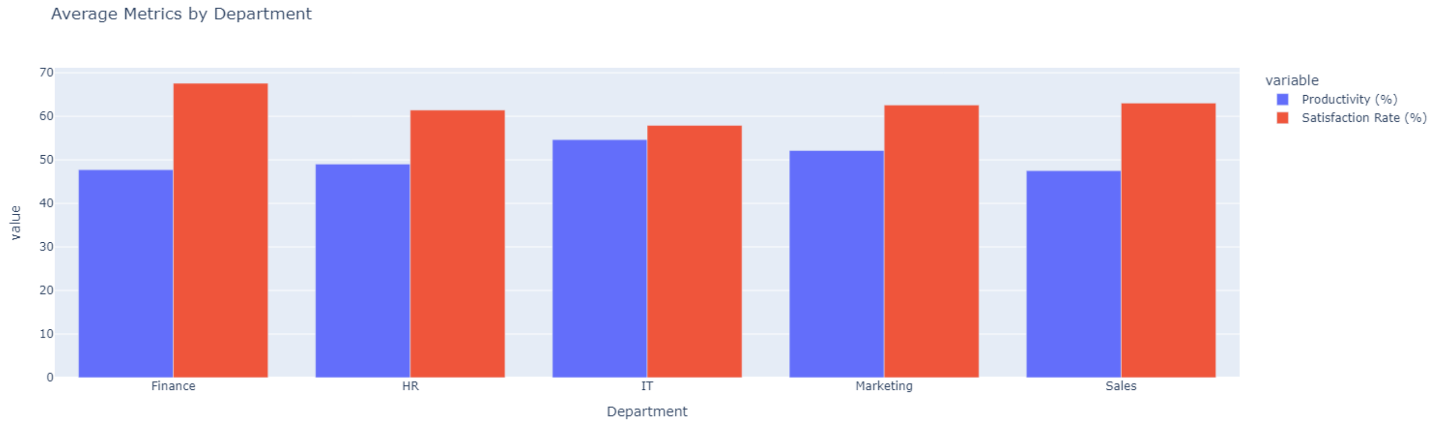


Fig 5.3.3 Average Productivity and Satisfaction rate by Department

5.4 Salary vs. Performance/Satisfaction

The Salary vs. Performance/Satisfaction analysis explores the relationship between employee salary and their productivity and satisfaction levels. It highlights how salary influences these metrics, identifying trends such as higher salaries correlating with improved productivity or satisfaction. The analysis also compares average metrics across different salary ranges for further insights.

1.Salary vs Productivity

2.Salary vs Satisfaction Rate

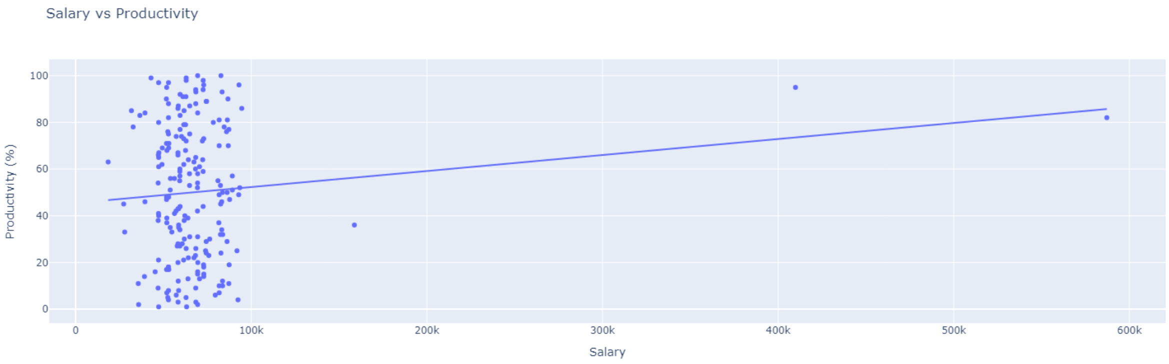


Fig 5.4.1 Salary vs Productivity

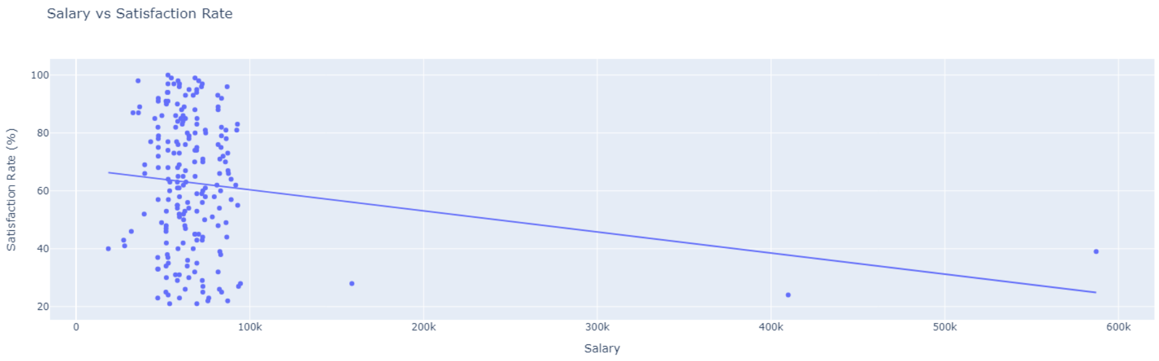
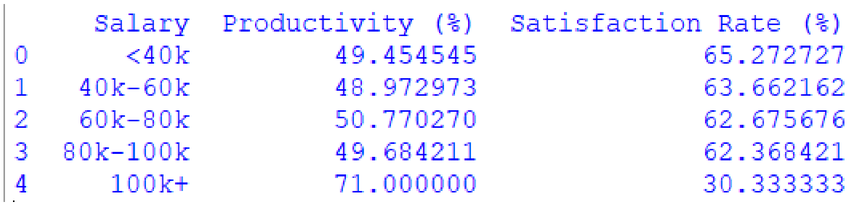


Fig 5.4.2 Salary vs Satisfaction Rate



```
     Salary  Productivity (%)  Satisfaction Rate (%)
0      <40k         49.454545              65.272727
1   40k-60k         48.972973              63.662162
2   60k-80k         50.770270              62.675676
3  80k-100k         49.684211              62.368421
4     100k+         71.000000              30.333333
```

Fig 5.4.3 Productivity and Satisfaction Rate for Salary

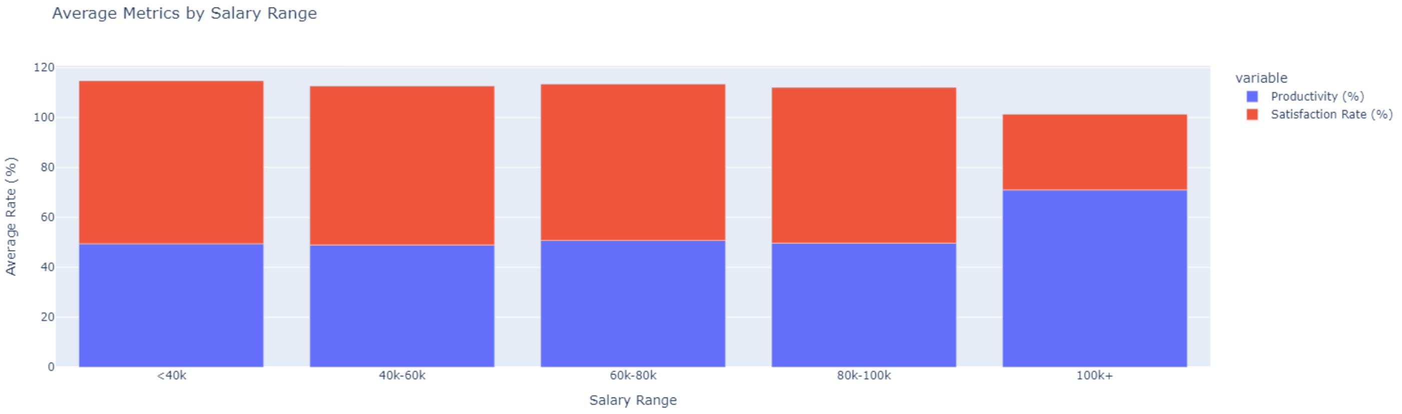


Fig 5.4.4 Average Metrics by Salary Range

5.5 Feedback Analysis

Feedback Analysis focuses on understanding employee feedback patterns. The distribution of feedback scores is analyzed to see how frequently different scores occur. Correlation analysis is then performed to determine how feedback scores relate to productivity and satisfaction rates, offering insights into how feedback impacts employee performance and well-being.

1. Feedback Scores Distribution
2. Correlation between Feedback Scores and Other Metrics

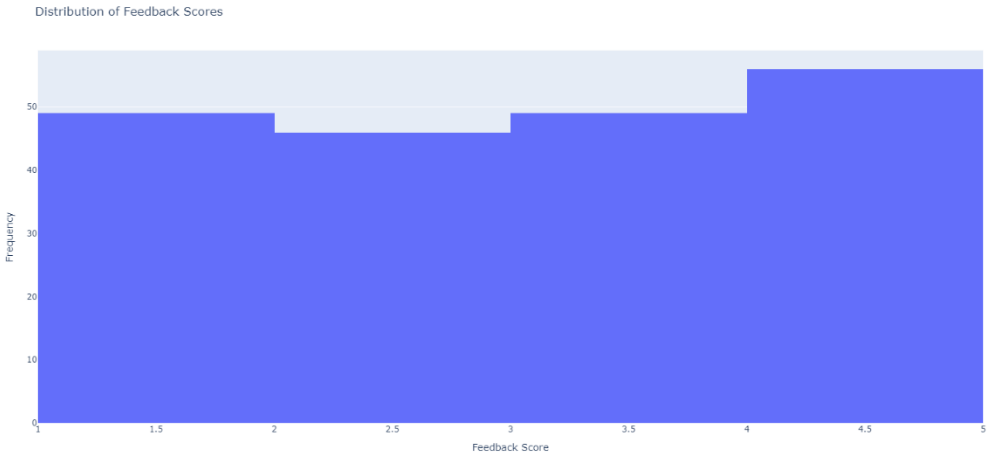


Fig 5.5.1 Distribution of Feedback Scores

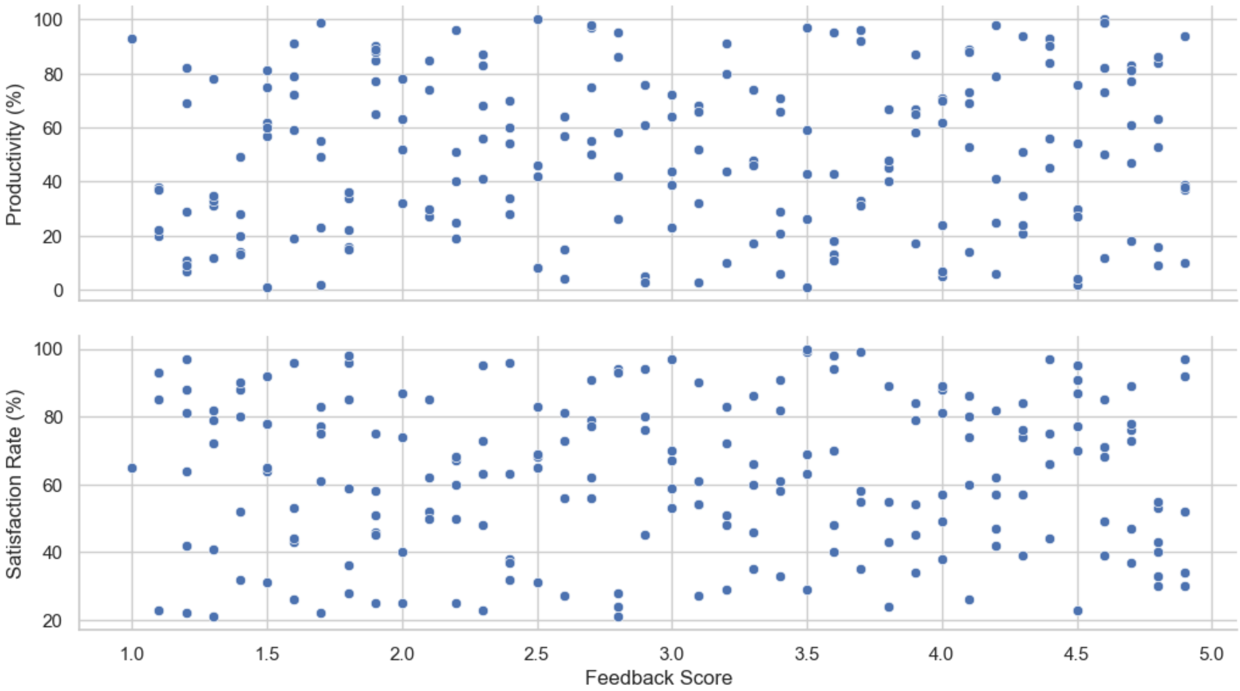


Fig 5.5.2 Correlation between Feedback Scores and Other Metrics

CHAPTER-6
CONCLUSION

The analysis of the employee performance and satisfaction dataset has provided valuable insights into how various factors impact productivity and overall employee satisfaction. By exploring demographic data, we observed that age and gender distributions have subtle but notable effects on productivity and satisfaction rates. Additionally, a correlation analysis revealed that the number of projects completed, along with feedback scores, plays a crucial role in influencing both productivity and satisfaction levels.

The project has also highlighted how salary variations affect employee performance. Higher salary ranges generally correspond to increased productivity and satisfaction, although the relationship is not always linear. Through the regression model and feedback analysis, we identified key areas where employee engagement can be improved, particularly for those with lower feedback scores or those handling heavier workloads.

By incorporating visualizations, the project provided HR professionals with an interactive tool to explore these insights, making it easier to spot trends and patterns in the data. This solution not only offers a deeper understanding of employee behaviour but also serves as a foundation for creating actionable strategies aimed at improving both individual performance and overall organizational efficiency.

The results reinforce the importance of data-driven decision-making in enhancing employee satisfaction and productivity.

CHAPTER-7
FUTURE ENHANCEMENTS

Future enhancements for this project could include integrating more advanced machine learning algorithms, such as decision trees or clustering, to uncover deeper patterns in employee performance and satisfaction. Expanding the dataset to include additional variables like work hours, training opportunities, or employee tenure could provide more comprehensive insights. Implementing real-time data updates and incorporating interactive features like predictive analytics and trend forecasting would further enhance the dashboard’s utility. Additionally, integrating natural language processing for more nuanced sentiment analysis of feedback can offer richer insights into employee sentiments and areas for improvement.

REFERENCES

1.C.R. Kothari, *Research Methodology: Methods and Techniques*. 2nd ed. New Age International Publishers, 2004.

2. J. Brownlee, *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End*. Machine Learning Mastery, 2016.

3. P. Alpaydin, *Introduction to Machine Learning*. 4th ed. MIT Press, 2020.

4. T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

5. H. Wickham and R. Grolemund, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2016.

6. Plotly, *Plotly Express Documentation*. [Online]. Available: https://plotly.com/python/plotly-express/

7.Scikit-learn, *Scikit-learn Documentation*. [Online]. Available: https://scikit-learn.org/stable/user_guide.html

8. S. E. Fienberg, *The Analysis of Cross-Classified Categorical Data*. MIT Press, 1970.

APPENDIX

SOURCE CODE

```
import pandas as pd
import numpy as np

#Library for visualization
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from wordcloud import WordCloud
from plotly.offline import iplot
from plotly.subplots import make_subplots
import missingno as msno
from pandas.plotting import parallel_coordinates
import altair as alt

#Library for building machine learning models
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report, accuracy_score

# Full path to the CSV file
file_path = r'C:\Users\mario\Desktop\hr_dashboard_data.csv'  # Use raw string by prefixing with r

# Read the CSV file
df = pd.read_csv(file_path)

# Display the first few rows of the dataframe
print(df.head())

print('Number of rows:', df.shape[0])
print('Number of columns:', df.shape[1])

duplicated_data = df.duplicated().any()
print(duplicated_data)

df.describe()
print(df.describe())

df.isnull().sum()
print(df.isnull().sum())
```

```
  df.dtypes
  print(df.dtypes)

  df.hist(figsize=(18,10))
  plt.show()

#Age Distribution
histogram_trace = go.Histogram(
    x=df['Age'],
    nbinsx=10,  # Number of bins
    histnorm='percent',  # Normalize histogram to percentage
    name='Age Distribution'
)

# Create the figure and add the trace
fig = make_subplots(rows=1, cols=1)
fig.add_trace(histogram_trace)

# Update layout of the plot
fig.update_layout(
    title='Age Distribution of Employees',
    xaxis_title='Age',
    yaxis_title='Percent',
    template='plotly',
    width=800,
    height=500
)

# Display the figure
fig.show()

#Gender Distribution
# Create a copy of the DataFrame and count gender distribution
gender_plot = df['Gender'].value_counts()

# Create a subplot for the pie chart
fig = make_subplots(rows=1, cols=1, specs=[[{'type': 'pie'}]], subplot_titles=['Gender Distribution'])

# Add the pie chart trace
fig.add_trace(go.Pie(values=gender_plot.values, labels=gender_plot.index), row=1, col=1)

# Update the chart layout and design
fig.update_traces(hoverinfo='label', textfont_size=18, textposition='auto',
                  marker=dict(colors=["#9467BD", "#1F77B4"], line=dict(color='white', width=2)))

fig.update_layout(title="<b>Gender Distribution of the Employees</b>", title_x=0.5, title_y=0.95,
                  template='xgridoff', width=800, height=500)
```

```
# Display the figure
fig.show()



#Employee Distribution by Department and Position
fig = px.treemap(
    df,
    path=['Department','Position'],
    values=[1] * len(df),
    title='Employee Distribution by Department and Position'
)
fig.show()

#Average Productivity and Satisfaction rate by Department and Position
avg_metric_dept = df.groupby('Department')[['Productivity (%)', 'Satisfaction Rate
(%)']].mean().reset_index()
avg_metric_pos = df.groupby('Position')[['Productivity (%)', 'Satisfaction Rate (%)']].mean().reset_index()

fig_dept = px.bar(
    avg_metric_dept,
    x='Department',
    y=['Productivity (%)', 'Satisfaction Rate (%)'],
    title='Average Metrics by Department',
    labels={'Department': 'Department'},
    height=500
)
fig_dept.update_layout(barmode='group')

fig_pos = px.bar(
    avg_metric_pos,
    x='Position',
    y=['Productivity (%)', 'Satisfaction Rate (%)'],
    title='Average Metrics by Position',
    labels={'Position': 'Position'},
    height=500
)
fig_pos.update_layout(barmode='group')

fig_dept.show()
fig_pos.show()

#Salary vs. Performance/Satisfaction
# Scatter plot for Salary vs Productivity
fig_prod = px.scatter(
    df,
    x='Salary',
    y='Productivity (%)',
    title='Salary vs Productivity',
    trendline='ols',  # Adds a trendline using Ordinary Least Squares (OLS)
    labels={'Salary': 'Salary', 'Productivity (%)': 'Productivity (%)'},
    height=500
)


# Scatter plot for Salary vs Satisfaction Rate
fig_satis = px.scatter(
    df,
    x='Salary',
    y='Satisfaction Rate (%)',
    title='Salary vs Satisfaction Rate',
    trendline='ols',  # Adds a trendline using Ordinary Least Squares (OLS)
    labels={'Salary': 'Salary', 'Satisfaction Rate (%)': 'Satisfaction Rate (%)'},
    height=500
)

# Display the plots
fig_prod.show()
fig_satis.show()

salary_ranges = pd.cut(df['Salary'], bins=[0, 40000, 60000, 80000, 100000, float('inf')],
                       labels=['<40k', '40k-60k', '60k-80k', '80k-100k', '100k+'])
average_metrics_by_salary = df.groupby(salary_ranges)[['Productivity (%)', 'Satisfaction Rate (%)']].mean().reset_index()

print(average_metrics_by_salary)

fig_avg_metric = px.bar(
    average_metrics_by_salary,
    x='Salary',
    y=['Productivity (%)', 'Satisfaction Rate (%)'],
    title='Average Metrics by Salary Range',
    labels={'Salary': 'Salary Range', 'value': 'Average Rate (%)'},
    height=500
)

fig_avg_metric.show()

# Feedback Scores Distribution
fig = px.histogram(df, x='Feedback Score', nbins=5, title='Distribution of Feedback Scores')
fig.update_layout(xaxis_title='Feedback Score', yaxis_title='Frequency')
fig.show()
```

```
# Correlation between Feedback Scores and Other Metrics
sns.set(style="whitegrid")  # Set the style to whitegrid
g = sns.pairplot(df, x_vars=['Feedback Score'], y_vars=['Productivity (%)', 'Satisfaction Rate (%)'],
kind='scatter')

# Adjust the layout of the pairplot
g.fig.suptitle('Correlation between Feedback Scores and Metrics', y=1.02)  # Set the title above the plots
plt.tight_layout(rect=[0, 0.03, 1, 0.95])  # Adjust the layout to avoid title overlap
plt.show()
```