Data Exploration Using R

OVERVIEW

This project analyses public datasets using statistical and visualisation techniques in R, uncovering key patterns, relationships, and insights.

OBJECTIVES

1. Perform univariate and multivariate analyses.
2. Use regression models and correlation to study variable relationships.
3. Apply PCA for dimensionality reduction and variance analysis.

DATASETS

1. USArrests: Crime rates in the US.
2. AirQuality: Air quality in New York.
3. swiss: Socio-economic data of Swiss provinces.
4. mtcars: Vehicle performance metrics.

OUTCOME

The project provides insights into each dataset through robust analysis, supported by clear visuals and reproducible R code.

Submitted by: Prabuddha Durge

Roll No.: 24BM6JP17
Date: 22 Nov 2024

Github Link: https://github.com/prabuddhadurge/DataExplorationUsingR

INDEX

Exploration of the Swiss Dataset

Overview

The `swiss` dataset (47 observations, 6 numerical variables) captures socio-economic factors of Swiss provinces, including `Fertility`, `Agriculture`, `Education`, and `Infant.Mortality`.

Key Findings

Univariate Analysis

- **Fertility Summary**:
  Mean: **70.14**, Median: **70.40**, SD: **12.49**, Range: **35.00–92.50**.
  The distribution is slightly right-skewed with potential outliers above **90**.

Correlation Analysis

- **Fertility vs. Education**:
  Pearson correlation: **-0.664**. Higher education is moderately linked to lower fertility.

Regression Analysis

- Predicting `Fertility` using `Agriculture` and `Examination`:
  - **Intercept**: **94.61**.
  - **Examination**: Significant negative effect ($p<0.001$).
  - Model Fit: Adjusted R2: **40.7%**.

Principal Component Analysis (PCA)

- **Explained Variance**:
  PC1 and PC2 capture **73.1%** of the total variance.
  - **PC1**: Strongly influenced by `Fertility` and `Agriculture`.
  - **PC2**: Driven by `Education` and `Examination`.
- **Insights**: Clustering reveals traditional vs. modern socio-economic influences.

Conclusion

Swiss provinces show a clear socio-economic divide. Fertility rates are negatively linked to education and examination performance, while PCA highlights distinct groupings driven by traditional and modern factors.

Exploration of USArrests Dataset

Overview

The USArrests dataset (50 observations, 4 variables) analyses violent crime rates across US states:

- **Murder**, **Assault**, **UrbanPop**, and **Rape**.

Key Findings

1. **Univariate Analysis**
    - **Murder**: Mean: **7.79**, SD: **4.36**, Range: **0.80–17.40**.
    - Distribution is right-skewed, with potential outliers above **15**.
2. **Correlation Analysis**
    - **Murder vs. Assault**: Pearson correlation: **0.802**.
      Higher assault rates are strongly associated with higher murder rates.
3. **Regression Analysis**
    - **Model**: Predicting Murder using Assault and Rape.
        - **Assault**: Significant positive effect ($p<0.001$), coefficient: **0.04**.
        - **Rape**: Not significant.
    - Adjusted R2: **62.95%**, FF-statistic: $p<0.001$.
4. **Principal Component Analysis (PCA)**
    - **PC1**: Explains **62.01%** variance, driven by Murder, Assault, and Rape.
    - **PC2**: Adds **24.74%**, highlighting UrbanPop.
    - Biplot reveals clustering of high-crime states.

Conclusion

- Strong correlations exist between violent crimes, particularly Murder and Assault.
- Regression identifies Assault as the vital predictor of Murder.
- PCA highlights socio-demographic groupings among states.

mtcars Dataset Analysis

Overview

The mtcars dataset (32 observations, 11 variables) summarises car attributes, including mpg (miles per gallon), hp(horsepower), and wt (weight in 1000 lbs).

Key Findings

1. **Univariate Analysis**
   - **mpg**: Mean: **20.09**, SD: **6.03**, Range: **10.40–33.90**. Slightly right-skewed distribution.
2. **Correlation**
   - **mpg vs. hp**: Strong negative correlation (r=−0.776), indicating higher horsepower reduces fuel efficiency.
3. **Regression Analysis**
   - **Model**: Predicting mpg using hp and wt.
     - **hp**: Negative impact (−0.032 mpg/unit, p=0.001).
     - **wt**: Larger negative impact (−3.88 mpg/1000 lbs, p<0.001).
   - Adjusted R2: **81.48%**.
4. **Principal Component Analysis (PCA)**
   - **PC1**: Explains **60.08%** of the variance, driven by mpg, hp, and wt.
   - **PC2**: Adds **24.09%**, highlighting secondary features.

Conclusion

- **Key Drivers**: Weight and horsepower significantly reduce fuel efficiency.
- **PCA Insights**: Vehicle performance dominates variability.

Exploration of Air Quality Dataset

Overview

The `airquality` dataset (153 observations, 6 variables) contains daily air quality measurements in New York from May to September 1973, including `Ozone`, `Solar Radiation`, `Wind`, and `Temperature`.

Key Findings

1. **Univariate Analysis**
   - **Ozone**: Mean: **42.13**, SD: **32.99**, Range: **1.00–168.00**.
     Distribution is right-skewed with missing values (37 NAs).
2. **Correlation Analysis**
   - **Ozone vs. Wind**: Pearson correlation: **-0.602**.
     A negative correlation suggests that higher wind speeds tend to be associated with lower ozone levels.
3. **Regression Analysis**
   - **Model**: Predicting `Ozone` using `Solar.R` and `Wind`.
     - **Intercept**: **77.25**.
     - **Solar Radiation**: Significant positive effect (p=0.0002), coefficient: **0.1004**.
     - **Wind**: Significant negative effect (p<0.001), coefficient: **-5.40**.
   - **Model Fit**: Adjusted R2: **43.93%**, indicating moderate explanatory power.
   - **Residual Analysis**: Residuals show variability, ranging from **-45.65** to **85.24**.
4. **Principal Component Analysis (PCA)**
   PCA on `Ozone`, `Solar.R`, `Wind`, and `Temp`:
   - **PC1**: Explains **59.0%** variance, dominated by `Ozone` and `Solar.R`.
   - **PC2**: Adds **22.37%**, capturing wind and temperature patterns.
   - **Cumulative Variance**: The first two components explain **81.36%** of variability.
   - **Biplot**: Reveals a strong clustering of observations based on `Ozone` and `Solar.R`.

Conclusion

- **Ozone Levels**: Strongly influenced by solar radiation and wind, with significant regression effects.
- **PCA Insights**: Variability in the data is mainly driven by ozone levels and solar radiation, with wind and temperature contributing less.
- **Model Performance**: The regression model explains nearly 44% of the variance, with strong effects of both `Solar.R` and `Wind` on `Ozone`.

LEARNINGS

This assignment provided a comprehensive understanding of data exploration and analysis techniques using R programming. Key learnings include:

1. **Univariate Analysis**: Gained proficiency in summarising and interpreting key metrics such as mean, median, standard deviation, and identifying patterns like skewness and outliers through visualisations like histograms and boxplots.
2. **Correlation and Regression:** Understood the importance of correlation coefficients to assess relationships between variables and leveraged linear regression models to predict outcomes and evaluate the significance of predictors.
3. **Principal Component Analysis (PCA)**: Learned to reduce dimensionality, interpret explained variance, and visualise data patterns using biplots, gaining insights into dominant factors influencing the dataset.
4. **Data Preprocessing**: Developed skills to handle missing values, scale variables, and prepare data for advanced analysis, ensuring accuracy in results.
5. **Visualisation and Interpretation**: Enhanced the ability to create meaningful plots (e.g., scatter plots, biplots) to visualise data relationships and communicate findings effectively.

SUMMARY

This project explores four **datasets** using **statistical techniques** and **R programming** to uncover insights and patterns.

The **Swiss dataset** reveals that **fertility rates** are negatively correlated with **education** (**r = -0.664**), with **regression** and **PCA** showing **socio-economic divides** and **73.1% variance** explained by **Fertility**, **Agriculture**, and **Education**.

The **USArrests dataset** identifies a strong correlation between **Murder** and **Assault** (**r = 0.802**), with **Assault** being a key predictor of **Murder**, while **PCA** explains **62.01% variance** through **violent crime metrics**.

The **mtcars dataset highlights** that **weight** and **horsepower** significantly reduce **fuel efficiency (mpg)**, with **PCA** attributing **60.08% variance** to **vehicle performance factors**.

Lastly, the **Air Quality dataset** shows that **ozone levels** are influenced by **solar radiation** and **wind**, with **regression** and **PCA** explaining **81.36% variance**, emphasising **environmental interactions**.