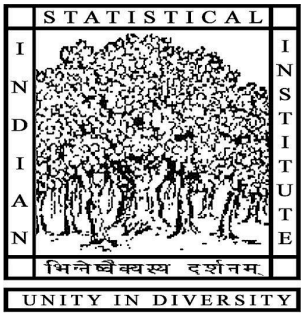
INDIAN STATISTICAL INSTITUTE, KOLKATA



POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS (Batch 10)

Subject: Statistical Structures in Data
Assignment

Name Moneesh B
Roll No. 24BM6JP34

Table of Contents:

Swiss Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The Swiss dataset provides standardized socio-economic and fertility measures across **47** Swiss provinces in 1888. It contains **six variables, all numerical:**

- **Fertility**: Standardized fertility measure.
- **Agriculture**: Percentage of males involved in agriculture.
- **Examination**: Percentage of draftees with the highest marks in army exams.
- **Education**: Percentage of draftees with education beyond primary school.
- **Catholic**: Percentage of Catholic population in the province.
- **Infant.Mortality**: Number of live births per 1,000 population.

Dataset Dimensions: **(47 rows and 6 columns)**.

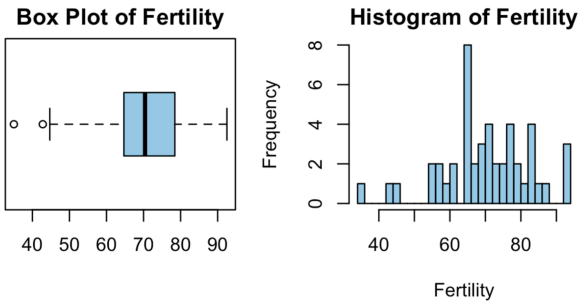
2. Summary Statistics

Summary statistics (Table 1) provide an overview of the central tendency and spread of each variable:

- **Fertility** has a moderate spread with a mean of 70.14.
- **Agriculture** varies widely, from 1.2% to 89.7%, indicating diverse agrarian influence.
- **Education** is positively skewed, with many provinces having low levels of education beyond primary school.
- **Catholic** shows high variability, with clustering near 0% and 100%, representing significant religious divides.

3. Distribution Visualization

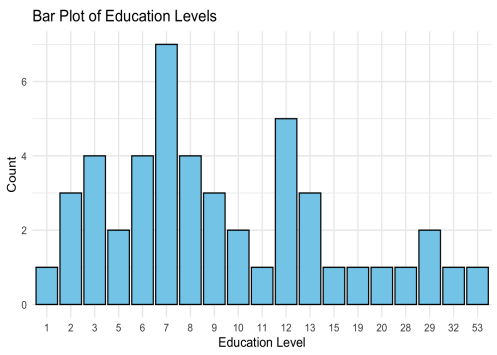
Variable Selected: Fertility



Box Plot Analysis: The box plot indicates that the Fertility values are moderately spread, with the central 50% (interquartile range) falling between approximately 62 and 78. The median Fertility is around 70, and a few outliers appear below 50, indicating provinces with notably lower Fertility rates compared to the rest.

Histogram Analysis: The histogram suggests that Fertility follows a **slightly right-skewed** distribution, with a concentration of provinces having Fertility values between 60 and 80.

4. Categorical Variable Analysis



Although the Swiss dataset contains no explicit categorical variables, the **Education** variable was treated as categorical for visualization purposes.

The bar plot shows clustering around 7% and 12%, indicating limited advanced education among draftees in most provinces.High education levels (above 30%) are rare, with a small number of provinces exceeding 50%.

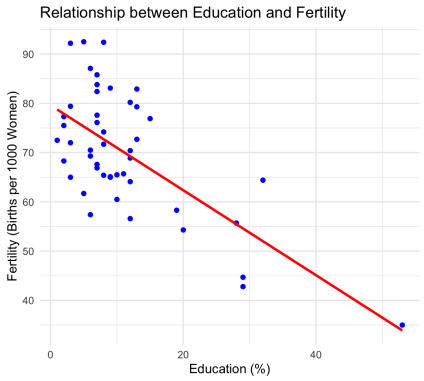
Education levels vary significantly, with most provinces reporting low percentages. This aligns with the time period and socio-economic conditions of 1888.

Multivariate Analysis Report

5. Correlation Analysis

- **Fertility and Education**: **Pearson Correlation Coefficient = -0.664**. This indicates a strong negative relationship, suggesting that higher education levels are associated with lower fertility rates.
- **Tabulation**: A correlation table (Table 2) represents these relationships.

6. Scatter Plot Visualization



- The scatter plot shows a clear negative trend between education levels and fertility.
- A linear regression trend line (in red) emphasizes this **inverse relationship**, corroborating the correlation coefficient of -0.664.
- Higher education often correlates with delayed childbearing and smaller family sizes, explaining this trend.

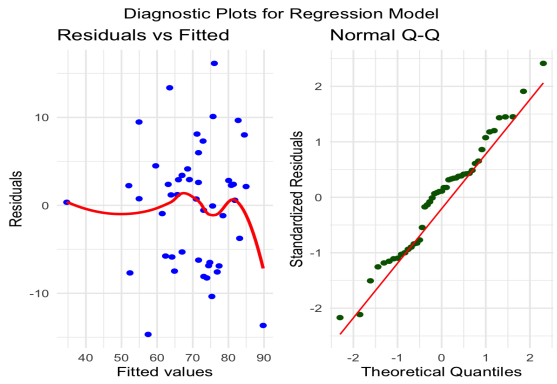
7. Multiple Regression

Fertility = **62.10 − 0.98**(Education) **− 0.15**(Agriculture) **+ 0.12**(Catholic) **+ 1.08**(Inf.Mortality)

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
```

- **Education**: Significant negative impact on fertility (-0.98, p < 0.001).
- **Agriculture**: Slight negative impact (-0.15, p < 0.05).
- **Catholic**: Positive influence (0.12, p < 0.001), reflecting cultural norms.
- **Infant Mortality**: Positive relationship (1.08, p < 0.01), indicating higher fertility rates in regions with higher infant mortality.
- **Adjusted R2=0.67**: 67% of the variability in fertility is explained by the model.
- **Residual Standard Error = 7.17:** Indicates moderate model accuracy.

8. Model Diagnostics



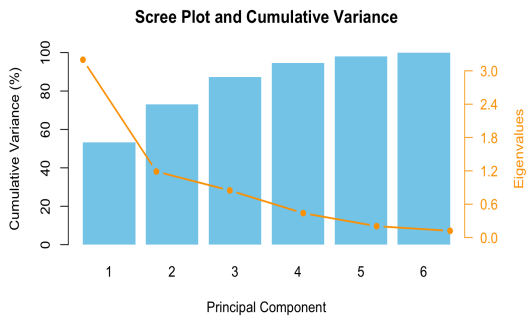
Residuals vs. Fitted: A slight curvature suggests some **non-linearity**, indicating potential model improvements.

Normal Q-Q Plot:Residuals follow the theoretical quantiles closely, confirming approximate normality.

The **Residuals vs. Fitted** plot does not show a clear pattern of increasing or decreasing spread of residuals, indicating

homoscedasticity (constant variance) is likely satisfied, The non-linearity in residuals suggests potential benefits from introducing interaction terms or non-linear transformations.

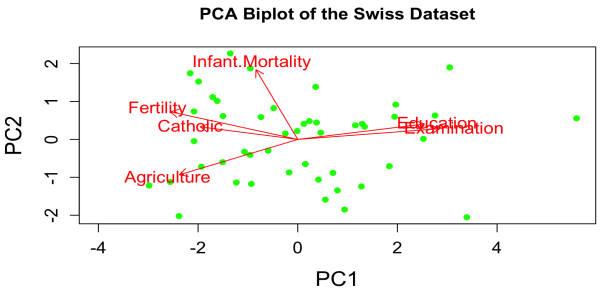
Advanced Analysis Report



9. Principal Component Analysis (PCA)

The scree plot demonstrates a sharp decline in variance after the second component, indicating that the first two components capture the majority of the variance: **PC1 explains 53.33% of the variance, PC2 explains 19.80% of the variance.** Based on the scree plot, retaining the first **two components(73% Explained Variance)** is sufficient for capturing the key dimensions of the dataset.

10. PCA Interpretation



Key Contributors to PC1: PC1 primarily contrasts socio-economic variables (**Examination** and **Education**) with traditional and religious aspects (**Fertility**, **Agriculture**, and **Catholic**). Provinces with higher socio-economic indicators align positively with PC1, while those with higher traditional indicators align negatively. (Loadings Table - (Table 3))

- **Key Contributors to PC2**: PC2 is predominantly driven by **Infant Mortality**, which has a high positive loading. **Agriculture** contributes negatively to PC2, indicating an inverse relationship with **Infant Mortality**.
- **Variable Relationships**: The directions of the arrows indicate correlations among variables. **Examination** and **Education** are closely aligned, showing a positive correlation. Similarly, **Fertility**, **Agriculture**, and **Catholic** are grouped together, indicating their association. **Infant Mortality** stands out with a unique influence on PC2.

Conclusion

The **Swiss dataset analysis** reveals significant socio-economic and demographic patterns. **Univariate analysis** highlights wide variability in **Fertility** and **Agriculture**, while **Education** shows low levels of advanced schooling. **Multivariate analysis** identifies a strong negative relationship between **Education** and **Fertility** (-0.664), with the regression model explaining 67% of the variability in fertility. Diagnostic checks confirm a reasonable model fit with minor non-linearity.

PCA shows that the first two components capture 73% of the variance. **PC1** contrasts socio-economic and traditional factors, while **PC2** highlights the impact of **Infant Mortality**, offering insights into the interplay of socio-economic, cultural, and demographic influences.

AirQuality Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The AirQuality dataset consists of six numerical variables representing meteorological and air pollution measures recorded in New York between May and September. The dataset contains **111 observations and 6 features** after removing rows with missing values.

2. Summary Statistics

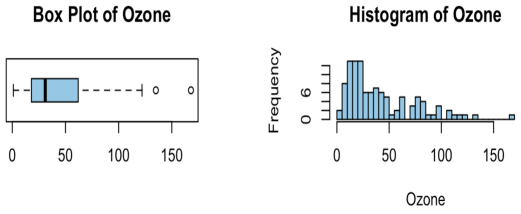
Summary statistics (Table 4) provide information on the central tendency and spread of variables:

- **Ozone** has a mean concentration of 42.1 ppb and exhibits right skewness with high outliers.
- **Solar.R** displays significant variability, with a mean of 184.8 langles.
- **Temp** shows a slight right skew, with a mean of 77.79°F.

Refer to **Table 2** in the appendix for detailed statistics of all variables.

3. Distribution Visualization

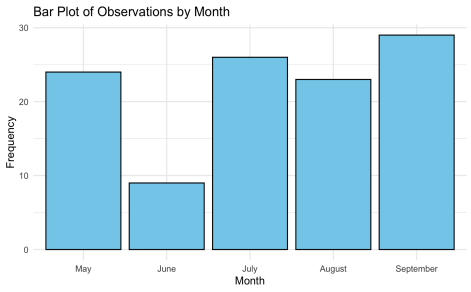
Variable Selected: Fertility



Box Plot: The Ozone data has **two clear outliers** above the upper whisker, with most values falling below 100. The distribution shows a moderate spread, and the median is approximately 30, indicating a concentration of lower ozone levels.

Histogram: The histogram shows a **right-skewed distribution**, with the majority of observations clustered between 0 and 50. Higher ozone levels (>100) are rare, and the frequency steadily declines as ozone levels increase.

4. Categorical Variable Analysis



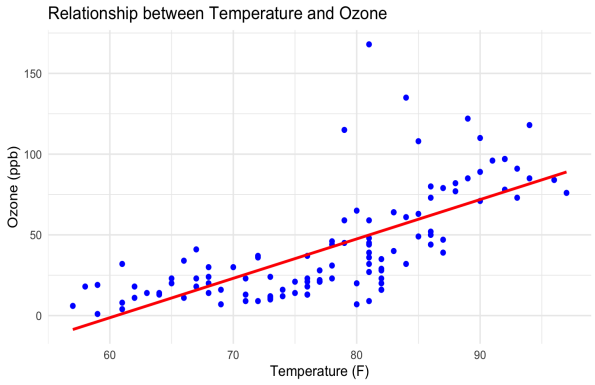
The **Month** variable is treated as categorical. The bar plot shows an uneven distribution of observations across months:

- Observations are highest in May and September, while June has the least coverage.
- The distribution ensures representation of seasonal patterns in air quality.

Multivariate Analysis Report

5. Correlation Analysis

- **Ozone and Temperature**: **Positive correlation (r=0.698)**, indicating higher temperatures are associated with elevated ozone levels. (Table 5)



6. Scatter Plot Visualization

The scatter plot shows a clear **positive linear relationship** between **temperature (°F)** and **ozone concentration (ppb)**, indicating that ozone levels tend to **increase with rising temperatures**. The **red regression line** highlights this trend, though some **variability** is observed, particularly at **higher temperatures**. This pattern suggests that **higher temperatures**, likely due to **increased photochemical activity**, contribute to **elevated ozone levels**, aligning with **environmental phenomena**.

7. Multiple Regression

Ozone = **−64.34 +1.65**(Temp) **+0.06**(Solar.R) **−3.33**(Wind).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Temp          1.65209    0.25353   6.516 2.42e-09 ***
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
```

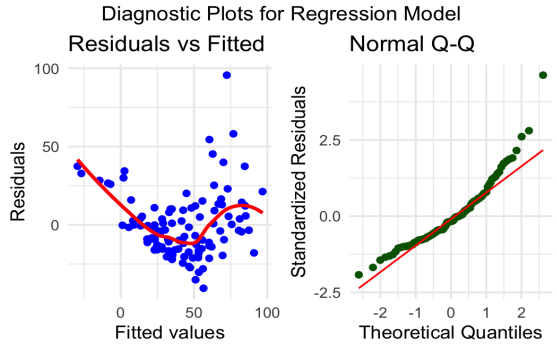
Temperature: Significant positive effect (p<0.001), suggesting that a 1°F increase in temperature raises ozone by 1.65 units.

Solar.R: Weak positive effect (p<0.05), indicating a marginal contribution to ozone levels.

Wind: Significant negative effect (p<0.001), showing dispersion effects reducing ozone concentrations.

Adjusted **R2=0.595**: The model explains ~60% of the variance in ozone levels. Residual standard error **(RSE=21.18)**: Indicates moderate accuracy.

8. Model Diagnostics

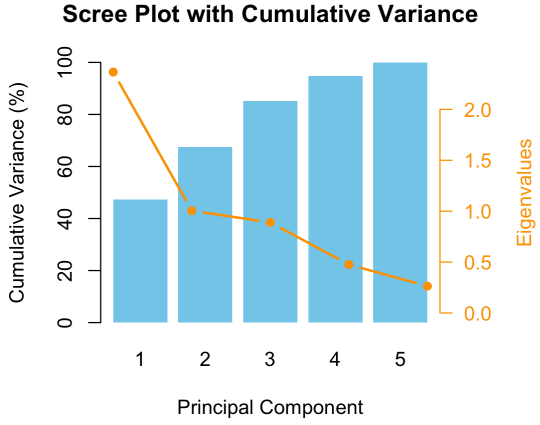


Residuals vs. Fitted: The plot shows a mild **non-linear pattern**, suggesting that the relationship between predictors and the response variable may not be entirely linear.

Normal Q-Q Plot: Residuals largely follow a weak normal distribution, with **deviations** at the tails.The model could benefit from non-linear or interaction terms.

Advanced Analysis Report

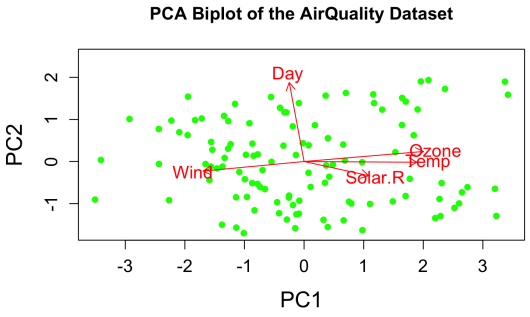
9. Principal Component Analysis (PCA)



Explained Variance: PC1 accounts for **47.34%** of the variance. PC2 adds **20.09%**. Together, the first two components explain **67.43%** of the total variance.

Scree Plot Analysis: The scree plot shows that the first two components capture the majority of the dataset's variability, justifying the inclusion of **two principal components**.

10. PCA Interpretation



PC1: Dominated by **Ozone** (0.586) and **Temp** (0.553) with high positive loadings, indicating a strong correlation between these variables. **Wind** (-0.496) contributes negatively, showing an inverse relationship with **Ozone** and **Temp**.

PC2: Heavily influenced by **Day** (0.970), which reflects seasonal or temporal effects. Other variables contribute minimally to this component.

Patterns and Groupings: **Ozone** and **Temp** are positively correlated, while **Wind** shows a negative association with them. The vertical alignment of **Day** indicates its variability is largely independent of other features, emphasizing its temporal significance. (Table 6)

Conclusion

The **AirQuality dataset analysis** highlights key meteorological and air pollution patterns. **Univariate analysis** shows significant variability in variables like **Ozone** and **Solar.R**, with **Ozone** displaying a right-skewed distribution and notable outliers. Seasonal patterns are captured through the **Month** variable, with uneven distribution across months ensuring seasonal representation.

Multivariate analysis reveals strong positive correlations between **Ozone** and **Temperature**, while **Wind** has a significant negative relationship with **Ozone**. The regression model explains ~60% of the variance in ozone levels but indicates a potential need for non-linear terms to improve fit. **PCA** simplifies the dataset by identifying two principal components, with the first capturing relationships among **Ozone**, **Temperature**, and **Wind**, and the second reflecting seasonal variations via **Day**.

Boston Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The Boston dataset contains **506 observations** and **14 variables**, providing socio-economic and housing-related metrics for Boston neighborhoods. Variables include numerical measures like crime rate, number of rooms per dwelling, and median home value, as well as categorical variables such as proximity to the Charles River. This diverse dataset captures key characteristics of urban housing conditions.

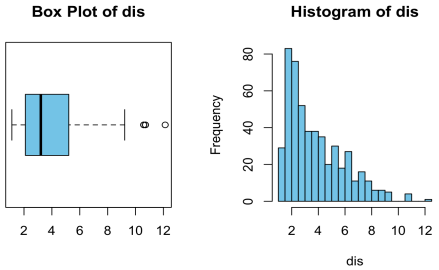
2. Summary Statistics

- **High Variability**: Numerical variables such as crime rate and property tax show wide ranges, reflecting the diversity in socio-economic conditions.
- **Key Trends**: Median home value and average room count are stable predictors of neighborhood quality.
- **Imbalances**: Categorical variables like proximity to the Charles River are heavily skewed, with a majority of neighborhoods not located near the river.

Refer to Table 7 for the complete summary statistics

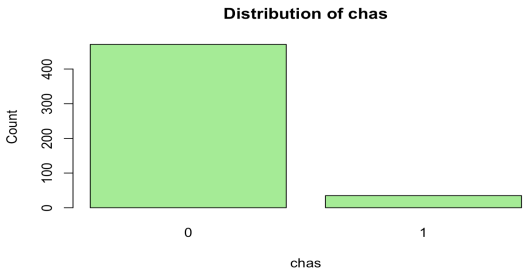
3. Distribution Visualization

Variable Selected: dis



Box Plot: The "dis" feature shows a **right-skewed distribution** with a median around 4.5 and a few **outliers** exceeding 10, representing unusually large distances to employment centers.
Histogram: The majority of observations are concentrated between **2 and 6**, with frequencies declining sharply as values increase, confirming the **right-skewed nature** of the data.

4. Categorical Variable Analysis



The categorical variable for proximity to the Charles River shows:

Imbalanced Distribution: A majority of neighborhoods (93%) are not near the river.

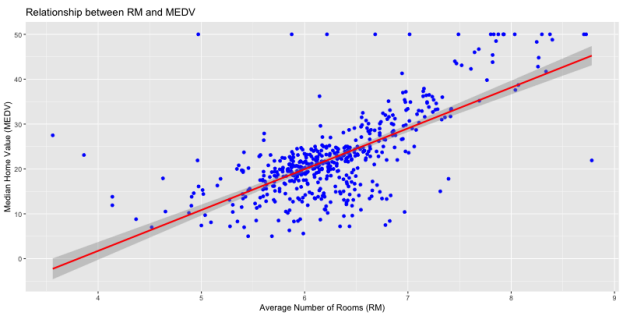
Insights: Limited observations near the river highlight spatial constraints in urban development.

Multivariate Analysis Report

5. Correlation Analysis

Strong Positive Correlation: The average number of rooms per dwelling (rm) is positively correlated with the median home value (medv) **(Pearson correlation coefficient = 0.695)**. (Table 8)

6. Scatter Plot Visualization



Positive Linear Relationship: As the average number of rooms (rm) increases, the median home value (medv) generally rises, indicating a direct relationship.

Trend Line: A linear regression trend line further emphasizes the positive association.

Variability: Data points spread around the trend line, with some higher room counts showing extreme median home values, potentially indicating luxury housing.

7. Multiple Regression

$$\textbf{medv} = -1.36 + 5.09\textbf{(rm)} - 0.64\textbf{(lstat)}$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.35827    3.17283  -0.428    0.669
rm           5.09479    0.44447  11.463   <2e-16 ***
lstat       -0.64236    0.04373 -14.689   <2e-16 ***
```

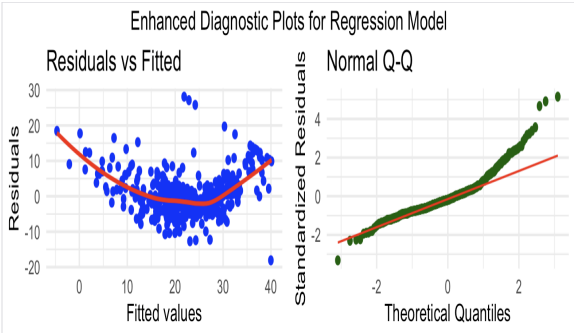
rm: For every additional room, the median home value increases by approximately $5,090, holding lstat constant.

lstat: For every 1% increase in the lower-status population, medv decreases by approximately $640.

Adjusted R2: 0.6371, indicating 63.7% of the variance in medv is explained by the model.

Significance: Both predictors (rm and lstat) are statistically significant with p-values < 0.001.

8. Model Diagnostics



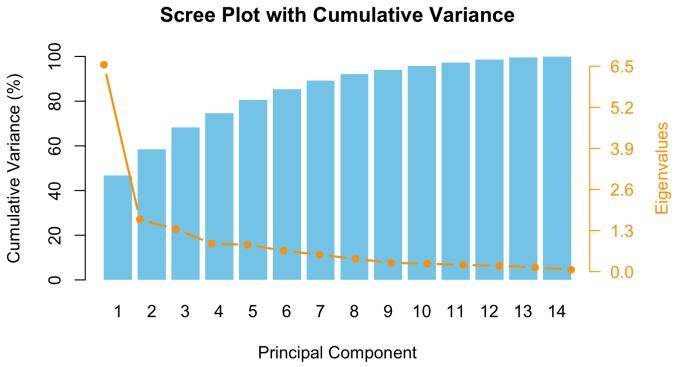
Diagnostic plots assess the assumptions of the regression model:

Residuals vs. Fitted Plot: The **U-shaped pattern** suggests **non-linearity**, indicating the model fails to fully capture the relationship between predictors and the response. However, there is no clear evidence of **heteroscedasticity**, therefore **homoscedasticity** holds.

Normal Q-Q Plot: Residuals largely follow a **normal distribution**, but deviations at the **tails** suggest the presence of outliers or non-normality in extreme values.

Advanced Analysis Report

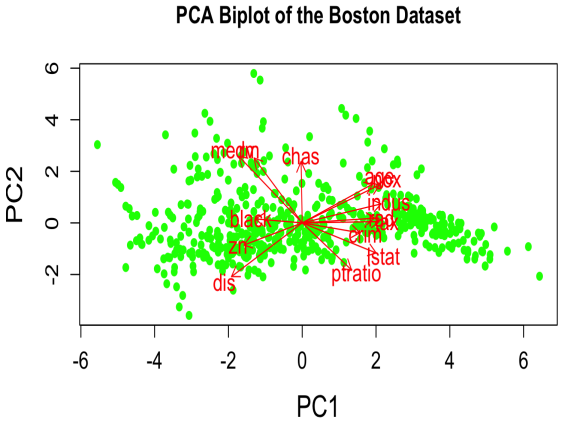
9. Principal Component Analysis (PCA)



Explained Variance: The first principal component (PC1) accounts for **46.8%** of the variance, while the **first two components** collectively explain **58.5%** of the variance. Adding a third component raises the cumulative variance explained to **68.1%**.

Components Selection: Based on the scree plot, which shows a steep decline in variance after PC2 and a more gradual decline thereafter, it is reasonable to reduce the dataset to two or three components without significant information loss. For interpretability, three components may be chosen to retain ~68% of the dataset's variability.

10. PCA Interpretation



PC1 (46.8% Variance): Dominated by variables such as rm (average number of rooms) and medv (median home value), with positive loadings. These variables indicate strong socio-economic characteristics tied to housing quality. Conversely, lstat (lower-status population) and ptratio (pupil-teacher ratio) load negatively on PC1, highlighting their inverse relationship with housing prices. (Table 9)

PC2 (11.8% Variance): Captures variations in urban and environmental factors, with strong positive contributions from nox (nitric oxides concentration) and indus (proportion of non-retail business acres). This component reflects industrial influence and environmental quality.

Patterns and Groupings:

- **Housing Quality Indicators:** Variables such as rm and medv cluster together, emphasizing their strong positive correlation.
- **Environmental Factors:** Variables like nox and indus align, reflecting their shared urban characteristics.
- **Inverse Relationships:** Features like lstat and ptratio align opposite to rm and medv, indicating their negative impact on housing prices.

The biplot thus reveals clear groupings of socio-economic, environmental, and urban characteristics, providing a comprehensive view of their contributions to housing price variability.

Conclusion

Univariate analysis shows **high variability** in metrics like **crime rate** and **property tax**, with **median home value** and **room count** as strong indicators of neighborhood quality.

Multivariate analysis reveals **housing prices** are positively influenced by **room count** but negatively impacted by **lower-status population**. **PCA** simplifies the dataset, showing that **socio-economic factors** and **urban characteristics** account for most of the variance.

Overall, the analysis emphasizes the significant role of **socio-economic** and **environmental factors** in shaping housing conditions.

CreditCard Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The CreditCard dataset consists of **1312 observations and 12 features**. It includes both numerical and categorical variables, representing consumer credit data such as income, expenditures, and ownership of credit cards. All rows are complete with no missing data.

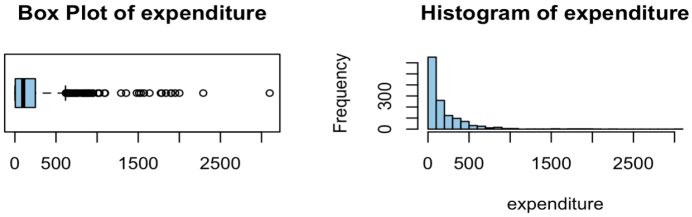
2. Summary Statistics

Summary statistics provide insights into the central tendency and spread of numerical variables: (Table 10)

- **Expenditure**: The average expenditure is 184.97 with a standard deviation of 272.71. The distribution is right-skewed with high outliers, ranging from 0 to 3099.51.
- **Income**: Average income is 3.37 with a standard deviation of 1.70. Income values range from 0.21 to 13.50.
- **Age**: Mean age is 33.38 years, ranging between 18 and 84 years.
- **Share**: Fraction of income spent has a mean of 0.0686, with a highly skewed distribution. Refer to Table 2 in the appendix for detailed statistics of all variables.

3. Distribution Visualization

Variable Selected: Expenditure

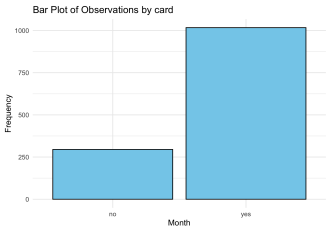


Box Plot: Shows significant outliers above the upper whisker. The median expenditure is approximately 101, with most values concentrated under 250.

Histogram: The distribution of expenditure is heavily **right-skewed**, with the majority of observations below 500.

4. Categorical Variable Analysis

Variable Selected: Card Ownership



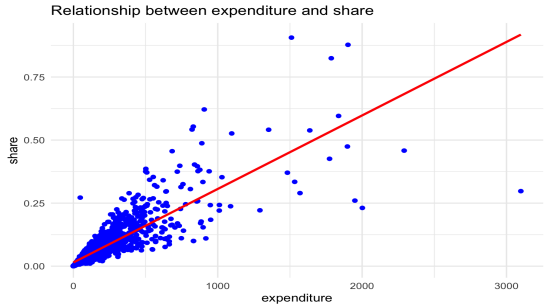
Bar Plot: The dataset is heavily imbalanced, with 1017 observations labeled as "yes" (cardholders) and 295 as "no" (non-cardholders). The imbalance highlights a potential challenge for predictive modeling, as models may disproportionately favor the majority class.

Multivariate Analysis Report

5. Correlation Analysis

Expenditure and Share: Strong positive correlation **(r = 0.839)**, indicating that higher expenditures are associated with a higher share of income spent. (Table 11)

6. Scatter Plot Visualization



Expenditure vs. Share: A strong linear relationship is observed, as higher expenditures correspond to a higher share of income spent. Notably, the variability in share is greater for lower expenditures, suggesting potential heteroscedasticity.

7. Multiple Regression

The multiple regression model predicts **expenditure** using **income**, **age**, and **share**:

Expenditure = − 161.33 + 52.63**(Income)** + 2463.72**(Share)**

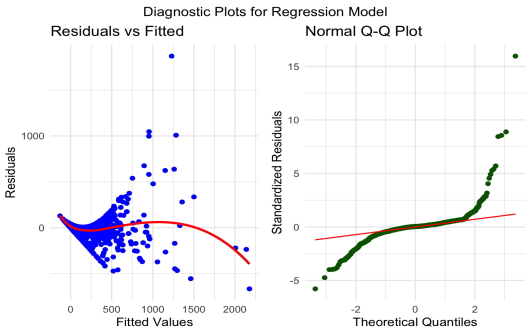
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -161.330      7.785  -20.72   <2e-16 ***
income        52.625      1.936   27.18   <2e-16 ***
share       2463.718     34.673   71.06   <2e-16 ***
```

Income: Retains a significant positive effect (**p < 0.001**), with a 1-unit increase in income leading to an increase of 52.63 units in expenditure.

Share: Continues to have a strong positive effect (**p < 0.001**), with a 1-unit increase in share associated with a substantial 2463.72-unit increase in expenditure.

The adjusted R2 is 0.81, the model explains ~81% of the variance in expenditure without a loss in predictive power. RSE is **118.8**.

8. Model Diagnostics

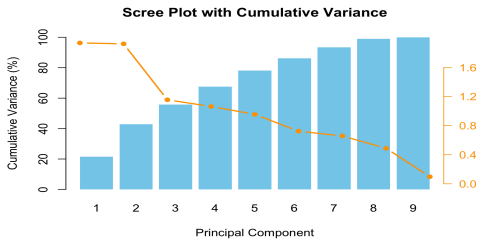


The **Residuals vs. Fitted** plot shows a clear non-linear pattern, indicating that the relationship between the predictors and the response variable may not be fully captured by the linear model. Additionally, the spread of residuals increases with fitted values, suggesting possible heteroscedasticity.

The **Normal Q-Q Plot** reveals deviations from normality, particularly in the tails, indicating that the residuals are not perfectly normally distributed. These diagnostics suggest that the model might benefit from transformations, addition of interaction terms, or non-linear predictors to better fit the data.

Advanced Analysis Report

9. Principal Component Analysis (PCA)

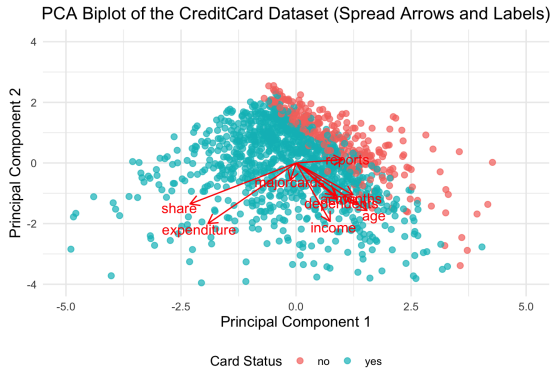


Explained Variance: The first two components explain ~43% of the total variance, while the first four components cumulatively capture ~67%, with diminishing contributions from subsequent components.

Component Selection: Based on the **elbow criterion**, the plot suggests that 2 components are sufficient, but including 4 components would capture a larger portion of the variability.

Conclusion: Selecting **2 components** balances simplicity and variance capture, though using **4 components** could be justified if a higher variance explanation (~67%) is desired.

10. PCA Interpretation



Loadings Interpretation: (Table 12)

PC1: Dominated by negative loadings from **share (-0.576)** and **expenditure (-0.478)**, indicating these variables strongly influence PC1 and are associated with spending behavior.

PC2: Influenced by **income (-0.482)** and **expenditure (-0.499)**, suggesting it captures demographic and financial factors.

Patterns and Groupings: Cardholders ("yes") align more with higher **expenditure** and **share**, clustering along PC1. Non-cardholders ("no") show greater variability along PC2, reflecting demographic differences in **income**, **age**, and **dependents**. Spending-related variables (**share**, **expenditure**) drive distinction along PC1, while **income** and **age** contribute to variance along PC2, separating groups by financial and demographic characteristics.

Conclusion

The CreditCard dataset analysis uncovers critical consumer spending patterns. **Univariate analysis** highlights significant variability in expenditure, income, and share, with expenditure exhibiting a right-skewed distribution and notable outliers. **Categorical analysis** reveals an imbalance in card ownership, with 77% of observations representing cardholders. **Multivariate analysis** shows strong positive relationships between expenditure, income, and share, with the regression model explaining ~81% of the variance in expenditure. However, diagnostics indicate potential improvements through non-linear or interaction terms. **PCA** reduces dimensionality effectively, with expenditure, share, and income contributing most to the first two components.

Appendix:

Swiss Data

Table 1: Summary Statistics

(return)

Variable	Min	Max	Mean	Median	Std. Dev
Fertility	35	92.5	70.14	70.4	12.49
Agriculture	1.2	89.7	50.66	54.1	22.71
Examination	3	37	16.49	16	7.98
Education	1	53	10.98	8	9.62
Catholic	2.15	100	41.14	15.14	41.7
Infant.Mortality	10.8	26.6	19.94	20	2.91

Table 2: Correlation Matrix

(return)

	Fertility	Agriculture	Examination	Education	Catholic	Infant Mortality
Fertility	1	0.353	-0.646	-0.664	0.464	0.417
Agriculture	0.353	1	-0.687	-0.64	0.401	-0.061
Examination	-0.646	-0.687	1	0.698	-0.573	-0.114
Education	-0.664	-0.64	0.698	1	-0.154	-0.099
Catholic	0.464	0.401	-0.573	-0.154	1	0.175
Infant Mortality	0.417	-0.061	-0.114	-0.099	0.175	1

Table 3: PCA Loading Factors

(return)

Variable	PC1	PC2
Fertility	-0.46	0.32
Agriculture	-0.42	-0.41
Examination	0.51	0.13
Education	0.45	0.18
Catholic	-0.35	0.15
Infant.Mortality	-0.15	0.81

AirQuality Data

Table 4: Summary Statistics Table

(return)

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std Dev
Ozone	1	18	31	42.1	62	168	33.28
Solar.R	7	113.5	207	184.8	255.5	334	91.15
Wind	2.3	7.4	9.7	9.94	11.5	20.7	3.56
Temp	57	71	79	77.79	84.5	97	9.53
Month	5	6	7	7.22	9	9	1.47
Day	1	9	16	15.95	22.5	31	8.71

Table 5: Correlation Matrix

(return)

	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	1	0.348	-0.612	0.699	-0.006	-0.005
Solar.R	0.348	1	-0.127	0.294	-0.058	-0.058
Wind	-0.612	-0.127	1	-0.497	0.05	0.05
Temp	0.699	0.294	-0.497	1	-0.097	-0.097
Month	-0.006	-0.058	0.05	-0.097	1	-0.179
Day	-0.005	-0.058	0.05	-0.097	-0.179	1

Table 6: Principal Component Loadings

(return)

Variable	PC1	PC2	PC3
Ozone	0.56	0.11	-0.19
Solar.R	0.52	-0.19	-0.33
Wind	-0.47	0.55	-0.23
Temp	0.46	-0.12	0.68
Month	-0.28	0.74	0.06
Day	-0.19	0.27	0.59

Boston Data

Table 7: Summary Statistics Table

(return)

Variable	Min	Max	Mean	Median	Std Dev
crim	0.01	88.98	3.61	0.26	8.60
zn	0.00	100.00	11.36	0.00	23.32
indus	0.46	27.74	11.14	9.69	6.86
chas	0.00	1.00	0.07	0.00	0.25
nox	0.39	0.87	0.55	0.54	0.12
rm	3.56	8.78	6.28	6.21	0.70
age	2.90	100.00	68.57	77.50	28.15
dis	1.13	12.13	3.80	3.21	2.11
rad	1.00	24.00	9.55	5.00	8.71
tax	187.00	711.00	408.24	330.00	168.54
ptratio	12.60	22.00	18.46	19.05	2.16
black	0.32	396.90	356.67	391.44	91.29
lstat	1.73	37.97	12.65	11.36	7.14
medv	5.00	50.00	22.53	21.20	9.20

Table 8: Correlation Matrix

(return)

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
crim	1.00	-0.20	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	-0.39	0.46	-0.39
zn	-0.20	1.00	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.32	-0.39	0.18	-0.41	0.36
indus	0.41	-0.53	1.00	0.06	0.76	-0.39	0.65	-0.71	0.60	0.72	0.38	-0.36	0.60	-0.48
chas	-0.06	-0.04	0.06	1.00	0.09	0.09	0.09	-0.10	-0.01	-0.04	-0.12	0.05	-0.05	0.18
nox	0.42	-0.52	0.76	0.09	1.00	-0.30	0.73	-0.77	0.61	0.67	0.19	-0.38	0.59	-0.43
rm	-0.22	0.31	-0.39	0.09	-0.30	1.00	-0.24	0.21	-0.21	-0.29	-0.36	0.13	-0.61	0.70
age	0.35	-0.57	0.65	0.09	0.73	-0.24	1.00	-0.75	0.46	0.51	0.26	-0.27	0.60	-0.38
dis	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75	1.00	-0.50	-0.53	-0.23	0.29	-0.50	0.25
rad	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.50	1.00	0.91	0.47	-0.44	0.49	-0.38
tax	0.58	-0.32	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1.00	0.46	-0.44	0.54	-0.47
ptratio	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.47	0.46	1.00	-0.18	0.37	-0.51
black	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1.00	-0.37	0.33
lstat	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	-0.37	1.00	-0.74
medv	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	1.00

Table 9: Principal Component Loadings

(return)

Variable	PC1	PC2
crim	0.24	-0.07
zn	-0.25	-0.15
indus	0.33	0.13
chas	-0.01	0.41
nox	0.33	0.25
rm	-0.20	0.43
age	0.30	0.26
dis	-0.30	-0.36
rad	0.30	0.03
tax	0.32	0.01
ptratio	0.21	-0.32
black	-0.20	0.03
lstat	0.31	-0.20
medv	-0.27	0.45

CreditCard Data

Table 10: Summary Statistics Table

(return)

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
reports	0.0	0.0	0.0	0.5	0.0	14.0
age	18.0	25.0	31.0	33.4	39.0	84.0
income	0.2	2.2	2.9	3.4	4.0	13.5
share	0.0	0.0	0.0	0.1	0.1	0.9
expenditure	0.0	4.6	101.2	185.0	249.0	3099.5
dependents	0.0	0.0	1.0	1.0	2.0	6.0
months	0.0	12.0	30.0	55.2	72.0	540.0
majorcards	0.0	1.0	1.0	0.8	1.0	1.0
active	0.0	2.0	6.0	7.0	11.0	46.0

Table 11: Correlation Matrix

(return)

Variable	reports	age	income	share	expenditure	dependents	months	majorcards	active
reports	1.00	0.04	0.01	-0.16	-0.14	0.02	0.05	-0.01	0.21
age	0.04	1.00	0.33	-0.12	0.02	0.22	0.45	0.01	0.19
income	0.01	0.33	1.00	-0.05	0.28	0.32	0.13	0.11	0.18
share	-0.16	-0.12	-0.05	1.00	0.84	-0.08	-0.05	0.05	-0.03
expenditure	-0.14	0.02	0.28	0.84	1.00	0.05	-0.03	0.08	0.05
dependents	0.02	0.22	0.32	-0.08	0.05	1.00	0.05	0.01	0.11
months	0.05	0.45	0.13	-0.05	-0.03	0.05	1.00	-0.04	0.10
majorcards	-0.01	0.01	0.11	0.05	0.08	0.01	-0.04	1.00	0.12
active	0.21	0.19	0.18	-0.03	0.05	0.11	0.10	0.12	1.00

Table 12: Principal Component Loadings

(return)

Variable	PC1	PC2
reports	0.25	0.03
age	0.39	-0.39
income	0.18	-0.48
share	-0.58	-0.34
expenditure	-0.48	-0.50
dependents	0.22	-0.30
months	0.31	-0.26
majorcards	-0.03	-0.14
active	0.21	-0.27