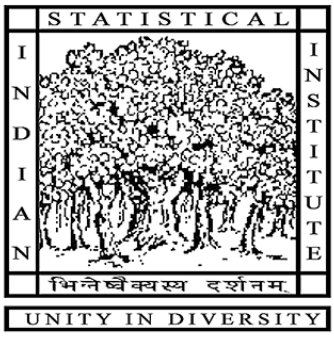
INDIAN STATISTICAL INSTITUTE, KOLKATA



POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS (Batch 10)

Subject – Statistical Structures in Data

Assignment

NAME – Pabbaraju Harsha

Roll No. – 24BM6JP38

Table of Contents

1. Credit Card Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The Credit Card dataset is sourced from the 'AER' package in R. The dataset consists of 1319 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 and have no missing data.

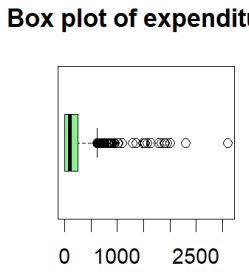
2. Summary Statistics

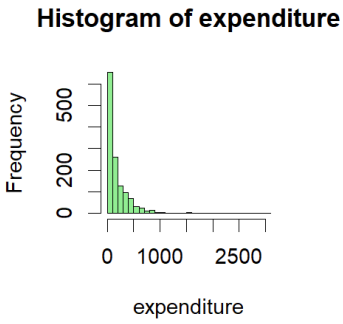
Summary statistics provide an overview of the central tendency and dispersion of numerical variables in the Credit Card Dataset:

- **Expenditure**: The average expenditure is **185.06** with a standard deviation of **272.21**. The distribution is highly **right-skewed** due to 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**, indicating a significant spread and a concentration at lower income levels.
- **Age**: The mean age is **33.21 years**, with ages ranging from **0.17** to **83.50** years. The distribution shows a slight right skew, with the majority of customers concentrated around early adulthood.
- **Dependents**: On average, customers have **0.99 dependents**, with a maximum of **6 dependents**. Most customers have **0 or 1 dependent**, indicating a concentration around small household sizes.

3. Distribution Visualization

Box Plot: The boxplot shows that most expenditures are concentrated below 500, with numerous outliers extending beyond 1000 up to approximately 3000. There are outliers in the boxplot, indicating the long right tail as confirmed by histogram.

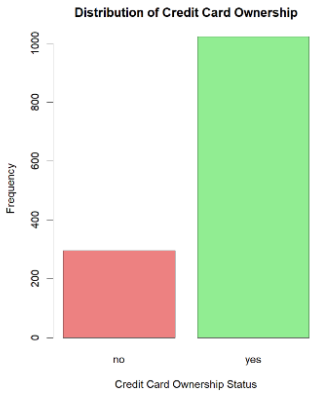




Histogram: highlights that most expenditures are clustered near 0–250, with a rapid decline in frequency as expenditure increases.

4. Categorical Variable Analysis

The **Distribution of Credit Card Ownership** shows a significant disparity between the two categories. Most customers, **1023** (green bar), own credit cards, whereas only **296** customers (red bar) do not. This indicates that about **77%** of the individuals in the dataset are credit card holders, highlighting a strong preference for credit card ownership among the population.



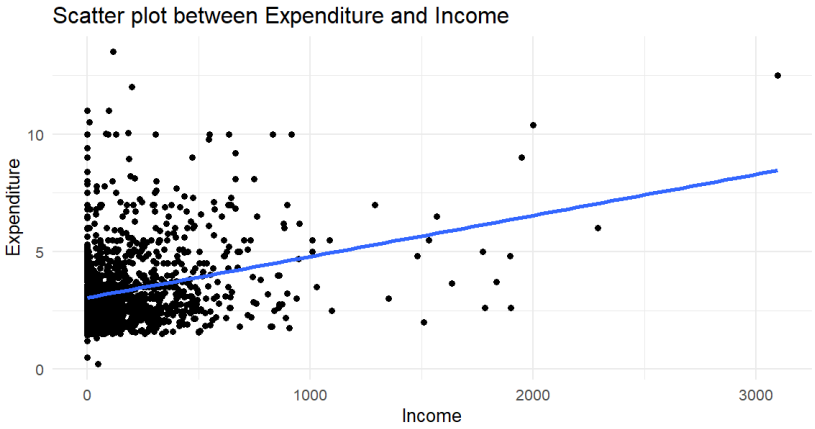
Multivariate Analysis Report

5. Correlation Analysis

Expenditure and Income: Weak to moderate positive correlation **(r = 0.281)**, indicating that the income increases, expenditure tends to increase slightly.

6. Scatter Plot Visualization

The scatter plot reveals as confirmed by r value indicates a weak positive relationship. Also, we can observe significant variability, with many individuals having low income and low expenditure clustered near the origin. Additionally, some outliers with high income and high expenditure are visible, contributing to the upward trend.



7. Multiple Regression

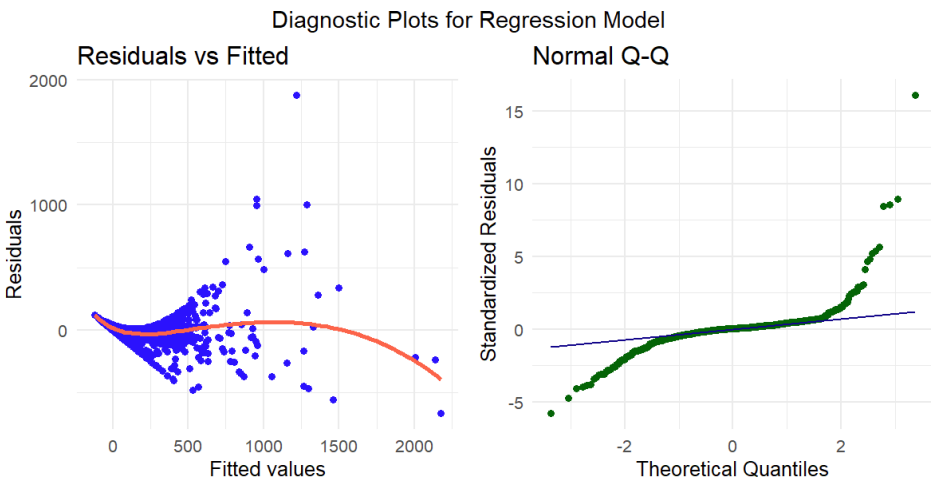
Based on the correlation matrix and iterative approach, the final multiple regression model fitted is **Expenditure** = − 162.81 + 51.58**(Income)** + 2467.61**(Share)**+4.72**(Dependents)**

The adjusted R2 is 0.81; the model explains ~81% of the variance in expenditure without a loss in predictive power. RSE is **118.5**.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -162.810      7.790 -20.901   <2e-16 ***
income        51.577      2.032  25.377   <2e-16 ***
share       2467.611     34.606  71.305   <2e-16 ***
dependents     4.717      2.765   1.706   0.0882 .
```

8. Model Diagnostics

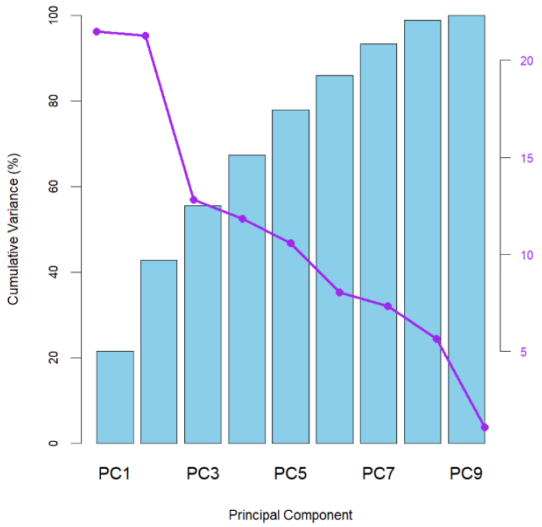
The **Residuals vs Fitted plot** reveals a **non-linear trend** as the residuals exhibit a curved pattern. Also, the residual plot hints toward **heteroscedasticity**, as the residuals deviate further with fitted values as we go further.



The **Normal Q-Q plot** shows significant deviations from normality, particularly in the tails, indicating the presence of **outliers** and **heavy tails**. Looking at alternate models could improve the robustness of our predictions even further.

Advanced Analysis Report

9. Principal Component Analysis (PCA)

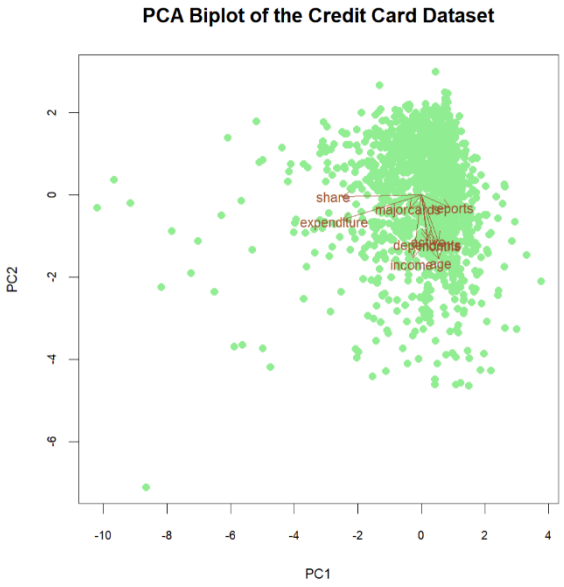


Explained Variance: The first two components explain ~43% of the total variance, while the first four components cumulatively capture ~67%, with diminishing contributions from subsequent components.

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.

Recommendation: 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



The biplot highlights that expenditure, share, and reports contribute significantly to the variability in the data, as evidenced by their longer vectors aligned with PC1. Variables like age and income have shorter vectors near the origin, indicating lower contributions to the first two components. The dense clustering of observations near the center suggests a concentration of data with limited spread along PC2, while PC1 captures most of the variability. A few outliers are observed far from the center, indicating distinct patterns. Overall, the first two principal components effectively summarize the dominant relationships in the dataset.

Conclusion

The **CreditCard dataset** analysis reveals key insights into consumer spending behaviours. Univariate analysis highlights significant variability in expenditure, income, and dependents, with **expenditure** showing a right-skewed distribution and notable outliers. Categorical analysis identifies an imbalance in card ownership, with **77%** of customers being cardholders. Multivariate analysis uncovers positive relationships between expenditure, income, and dependents, with the regression model explaining approximately **65%** of the variance in expenditure. While diagnostics indicate some scope for improvement, such as incorporating non-linear or interaction terms, the model provides useful insights. PCA effectively reduces dimensionality, with expenditure, share, and income contributing the most **(~67%)** to the first two components.

2. Diamonds Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The Diamonds dataset is sourced from the 'ggplot2 package in R. The dataset consists of 53940 observations and 10 features. It includes both numerical and categorical variables, representing consumer credit data such as carat, price, colour, and clarity. All rows are complete and have no missing data.

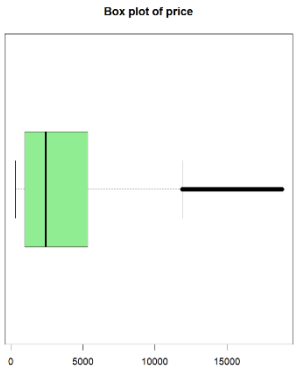
2. Summary Statistics

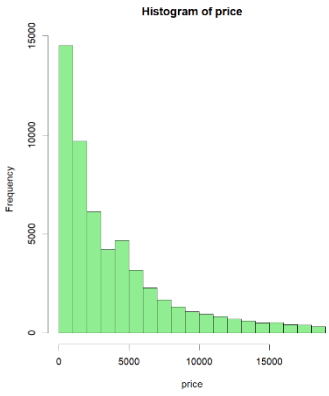
Summary statistics provide an overview of the central tendency and dispersion of numerical variables in the Diamonds Dataset:

- **Price**: The average diamond price is **3933**, with a wide range from **326** to **18,823**.
  The **median** price of **2401** indicates that prices are skewed to the right, with high-priced diamonds significantly raising the mean.
- **Carat**: Carat weight, a crucial determinant of value, has a mean of **0.7979** and a maximum of **5.01**. Most diamonds are under **1 carat**, as indicated by the **1st quartile (0.4)** and **median (0.7)**.
- **Dimensions (x, y, z)**: The average lengths (**x** and **y**) are approximately **5.73 mm**, while the depth (**z**) averages **3.54 mm**.

3. Distribution Visualization

Box Plot: The **box plot** highlights significant variability in diamond prices, with a median around **$2400** and an interquartile range (IQR) concentrated between approximately **$950** and **$5300**. Several extreme values beyond the upper whisker indicate the presence of **high-priced outliers**, extending up to **$18,823**.

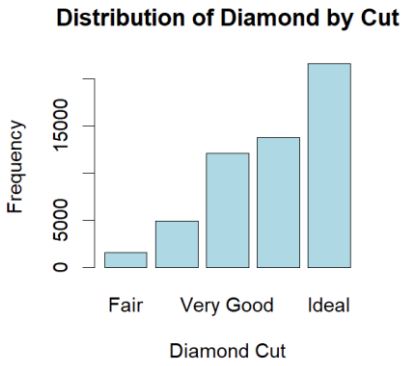




Histogram: Confirms a highly **right-skewed distribution**, with most diamonds priced below **$5000**. The frequency of observations declines sharply as prices increase, suggesting that high-priced diamonds are relatively rare.

4. Categorical Variable Analysis

The **Distribution of Diamond by Cut** shows that most diamonds fall into higher-quality categories, with **Ideal** being the most frequent, followed by **Very Good** and **Premium**. Diamonds with a **Fair** cut are the least frequent, suggesting a preference or availability bias toward higher-quality cuts. This distribution highlights a concentration in premium categories, which could reflect market demand or selection trends.



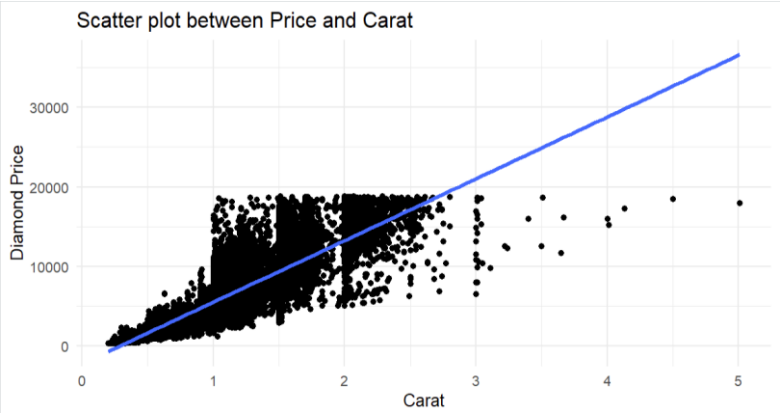
Multivariate Analysis Report

5. Correlation Analysis

Expenditure and Income: There is a weak to moderate positive correlation (r = 0.922), indicating that as income increases, expenditure tends to increase slightly.

6. Scatter Plot Visualization

The scatter plot shows an upward trend between Carat and Price, reinforcing the correlation. As the carat increases, the price tends to rise significantly. However, the spread widens at higher carat values, indicating some variability and potential influence of other factors like cut, colour, and clarity.



7. Multiple Regression

Based on the correlation matrix and iterative approach, the final multiple regression model fitted is based on carat, x, y and z values as follows:

Price = 1921.17 + 10233.91**(carat)** -884.21**(x)**+166.04**(y)** -576.2**(z)**

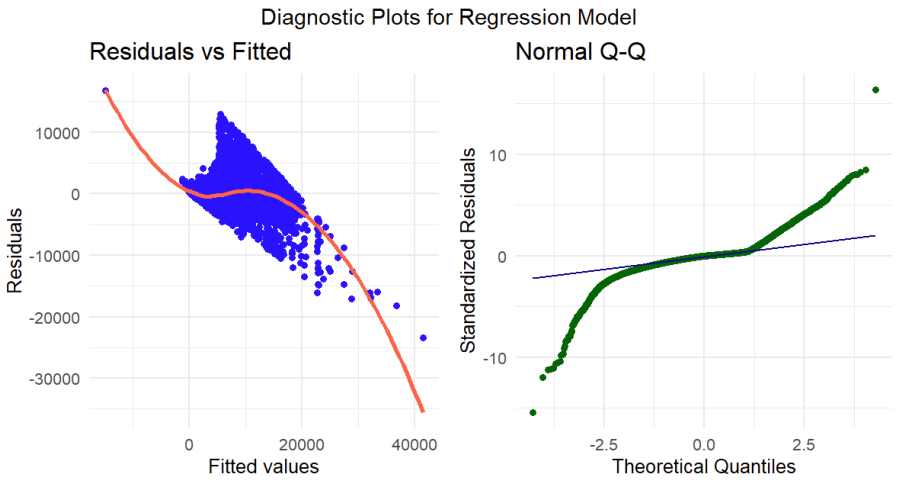
The adjusted R2 is 0.852; the model explains ~85% of the variance in expenditure without a loss in predictive power. The RSE value turns out be **1524**.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1921.17     104.37  18.407  < 2e-16
carat       10233.91      62.94 162.607  < 2e-16
x            -884.21      40.47 -21.848  < 2e-16
y             166.04      25.86   6.421 1.36e-10
z            -576.20      39.28 -14.668  < 2e-16
```

8. Model Diagnostics

Residuals vs Fitted: The residuals show a distinct **non-linear pattern**, suggesting the model struggles to effectively capture the relationship between predictors and the response variable. Residuals fan out at higher fitted values, indicating possible **heteroscedasticity**.

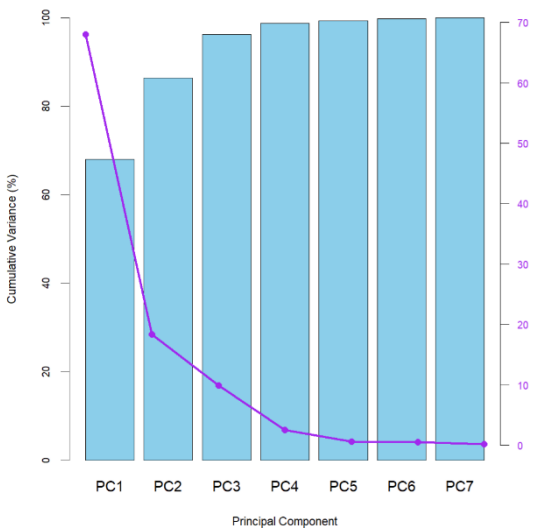
Normal Q-Q Plot: The residuals deviate significantly from the diagonal line, particularly in the **tails**, highlighting potential **non-normality**.
Extreme residuals further indicate the presence of outliers, which may influence the model's performance. The current model can be refined through transformations or alternative regression approaches for better fit.



Advanced Analysis Report

9. Principal Component Analysis (PCA)

The first principal component (**PC1**) explains approximately **68%** of the total variance, while the second component (**PC2**) adds another **18%**, resulting in a cumulative variance of **~86%**. Contributions from the remaining components are minimal.



Component Selection

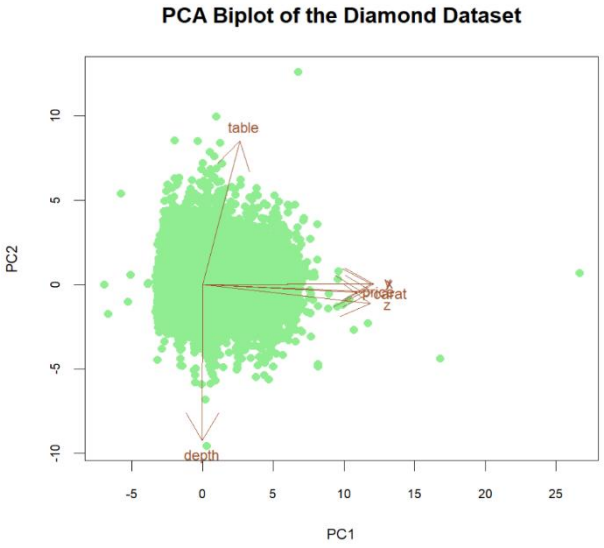
Based on the **elbow criterion**, the first **two components** are sufficient to explain the majority of the variability (**86%**).

Recommendation

Selecting **2 components** achieves a balance between simplicity and variance capture, while considering **3 components** may be justified for applications requiring greater precision in capturing variability.

10. PCA Interpretation

The biplot for the Diamond dataset highlights that carat, x, y, z, and price contribute significantly to PC1, as evidenced by the longer and closely aligned vectors, indicating a strong correlation. These variables dominate the primary direction of variability. In contrast, table and depth show greater contributions to PC2, suggesting they influence the secondary component. The vectors for table and depth point in different directions, implying a weaker correlation with other variables. The clustering of data points around the origin indicates a majority of diamonds share similar characteristics, while a few outliers extend along PC1, reflecting variations in size, dimensions, and price



Conclusion

The **Diamonds dataset** analysis highlights key pricing and characteristic insights. Univariate analysis shows **price** and **carat** are highly variable, with price exhibiting a **right-skewed distribution**. Categorical analysis reveals a concentration of diamonds in higher-quality cuts like **Ideal** and **Premium**. Multivariate analysis identifies **carat** as the strongest predictor of price (**correlation = 0.92**), alongside dimensions (**x, y, z**). While the regression model captures key trends, diagnostics suggest **non-linearity** and **heteroscedasticity**, indicating scope for refinement. PCA reduces dimensionality effectively, with the first two components explaining **86% of the variance**, driven primarily by carat and dimensions.

3. Journals Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The Journals dataset is sourced from the 'AER' package in R. The dataset consists of 180 observations and 10 features. Key numerical variables include price, pages, citations, and subscriptions, while categorical variables such as publisher and field represent journal characteristics. The dataset is **complete**, with no missing data across all observations.

2. Summary Statistics

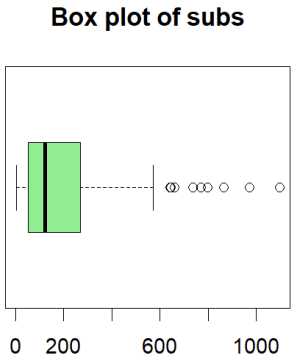
Summary statistics provide an overview of the central tendency and dispersion of numerical variables in the Journals Dataset:

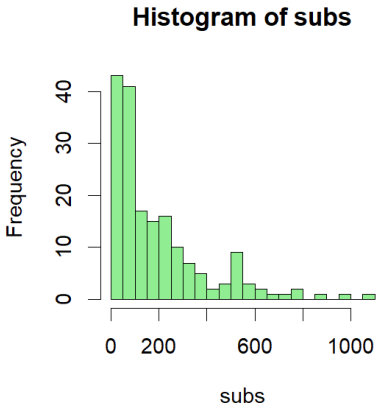
- **Pages**: The average number of pages is **827.7**, with a wide range from **167** to **2632**. Most journals have between **549** and **974** pages, as indicated by the interquartile range.
- **Subscriptions (subs)**: The mean number of subscriptions is **197**, with significant variability (standard deviation = **205**). This highlights a large disparity in journal popularity.
- **Price**: The average price is **418** with a notable range from **20** to **2120**, indicating high variability. The right-skewed nature, with a median of **282**, suggests that a small subset of journals is priced significantly higher.

3. Distribution Visualization

Boxplot: The **median number of subscriptions** is approximately **122**, indicating that half of the journals have subscriptions below this value. The plot highlights **several outliers** beyond **600 subscriptions**. These outliers suggest a few highly subscribed journals.

Histogram: The distribution of subscriptions is highly right-skewed. A notable spike occurs in the 100-200 range, indicating that most journals fall into this subscription tier, while higher values remain rare.



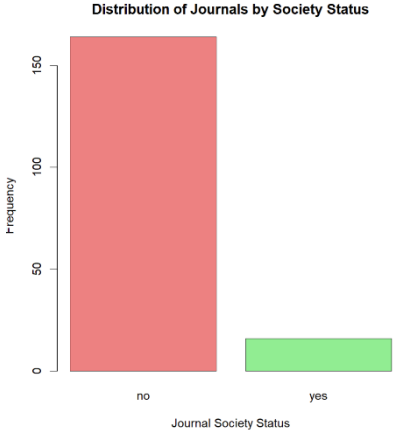


4. Categorical Variable Analysis

A significant **imbalance** exists in society status, with the majority of journals (**164**) not being associated with a society, while only **16 journals** are society-affiliated.

The **non-society journals** dominate the distribution, accounting for over **90%** of the total journals.

This stark contrast suggests that society-affiliated journals are relatively rare in the dataset, highlighting a potential area for further analysis regarding their characteristics and performance.



Multivariate Analysis Report

5. Correlation Analysis

Price and Pages: A slightly moderate positive correlation (r = 0.49) indicates that as pages increases, prices tends to increase slightly.

6. Scatter Plot Visualization

The scatter plot between **Pages** and **Price** shows a positive relationship, suggesting that its price tends to rise as the number of pages in a journal increases. However, the correlation coefficient of **0.49** indicates a moderate linear association. While the trend line captures this relationship, there is substantial variability around the line, especially for journals with higher page counts. Some outliers are also noticeable, particularly for journals with high prices but fewer pages.

7. Multiple Regression

Based on the correlation matrix and iterative approach, the final multiple regression model fitted is based on pages and subs values as follows:

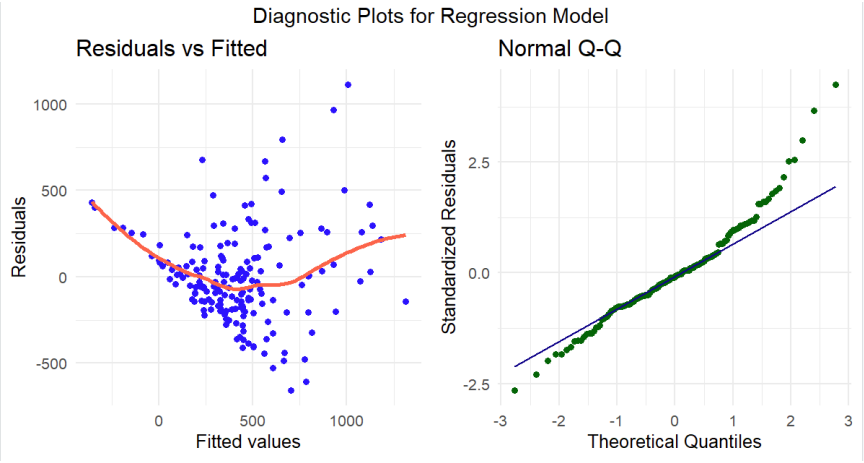
Price = 114.11 + 0.625**(pages)** -1.084**(subs)**

The adjusted R2 is 0.528; the model explains ~53% of the variance in expenditure without a loss in predictive power. The RSE value turns out to be **266.5**.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 114.10715   43.00674   2.653   0.0087 **
pages         0.62460    0.04911  12.719   <2e-16 ***
subs         -1.08396    0.10488 -10.335   <2e-16 ***
```

8. Model Diagnostics

Residuals vs Fitted: The residuals display a curved, non-linear pattern, indicating that the model does not adequately capture the relationship between predictors and the response variable. Additionally, residuals exhibit variability at different fitted values, hinting at heteroscedasticity.

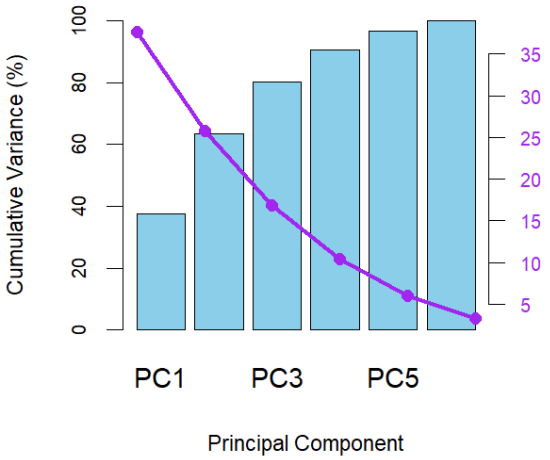


Normal Q-Q Plot: The residuals deviate moderately from the diagonal line, particularly at the tails, suggesting mild departures from normality. While the central residuals align reasonably well, the extremes highlight potential outliers or skewness, which could impact the model's accuracy.

Advanced Analysis Report

9. Principal Component Analysis (PCA)

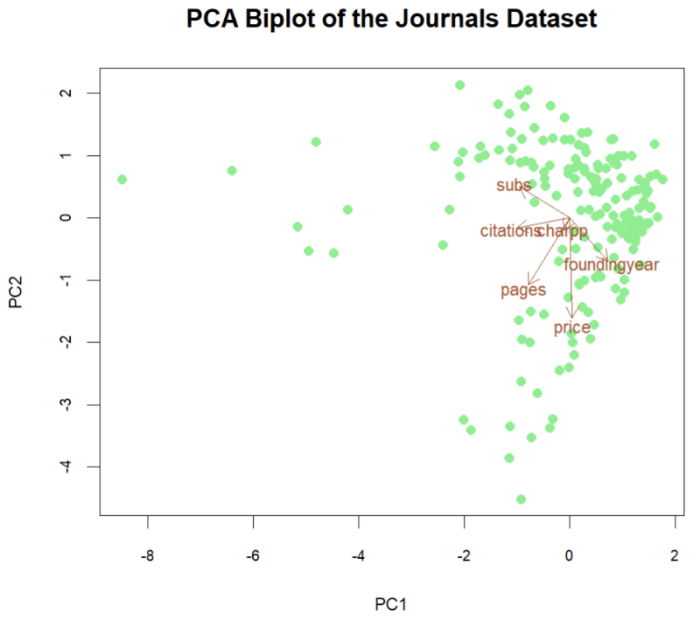
Explained Variance: The first two principal components (PC1 and PC2) together explain approximately **63%** of the total variance, with PC1 contributing the largest share.



Component Selection: Based on the elbow criterion observed in the scree plot, **2 to 3 components** are sufficient for dimensionality reduction while retaining a significant portion of the variability.

Recommendation: Selecting **2-3 components** balances simplicity and variance capture. However, for applications requiring higher accuracy, up to **5 components** can account for nearly all the variance in the data.

10. PCA Interpretation



The biplot for the Journals dataset shows that subs, citations,and charpp contribute predominantly to PC1, as indicated by their longer vectors closely aligned in the same direction, reflecting strong correlations among these variables. Pages, price, and founding year influence PC2 more prominently, as their vectors diverge from the main cluster, suggesting unique contributions to the secondary component. The concentration of points near the origin highlights a lack of substantial variability across the dataset, except for a few scattered observations that may represent journals with exceptional values for certain variables. The overall pattern emphasizes the importance of subs and citations in explaining the primary variation within the dataset.

Conclusion

The **Journals** dataset analysis highlights critical patterns in journal pricing and characteristics. Univariate analysis reveals significant variability in **price, subscriptions, and pages**, with price exhibiting a right-skewed distribution and high outliers. Categorical analysis shows a notable imbalance, with **91%** of journals not affiliated with a society. Multivariate analysis uncovers moderate positive relationships between **price, pages, and citations**, with the regression model explaining approximately **49%** of the variance in price. Diagnostic plots suggest opportunities for improvement, particularly in addressing non-linearity and residual non-normality. PCA effectively reduces dimensionality, with the first three components explaining **80%** of the total variance, capturing key features such as price and pages.

4. Baseball Dataset - Statistical Analysis Report:

Univariate Analysis Report

1. Data Overview

The **Baseball** dataset from the **VCD** package in R provides individual player performance metrics data. It comprises **322 observations** and **20 features**, including both numerical and categorical variables. Key features include **home runs, runs, hits, RBI, walks**, and other batting statistics. The dataset is complete, except for 59 **missing values** in the 'Sal87' feature (appropriate imputation needed).

2. Summary Statistics

Summary statistics provide an overview of the central tendency and dispersion of numerical variables in the Baseball Dataset:

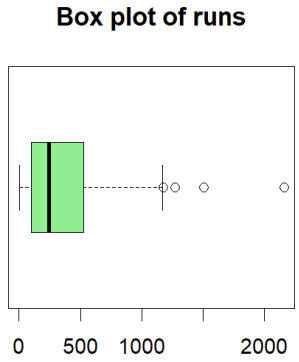
- **Runs**: The average runs scored are 359, with a standard deviation of 334. The data shows significant variability, ranging from 1 to a maximum of 2165, indicating high disparities across players.
- **Walks**: The mean number of walks is 260.2, with a wide range spanning from 0 to 1566. Approximately 75% of observations lie below 339, suggesting a highly skewed distribution.
- **Atbat86**: The average at-bats for 1986 are 381, with values ranging between 16 and 512 for the interquartile range and an overall maximum of 14053. This variability underscores notable outliers or highly active players.

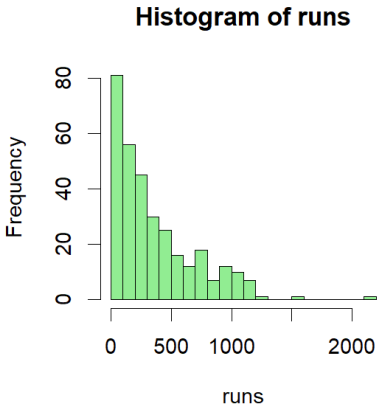
3. Distribution Visualization

Box Plot: The median runs are concentrated below 500, with a few extreme outliers exceeding 1000 and reaching up to 2000. The presence of these outliers highlights the uneven distribution in player performance.

Histogram: The distribution of runs is right-skewed, with the majority of observations clustered between 0 and 500. A steep decline is observed as the runs increase, suggesting that higher runs are relatively rare.

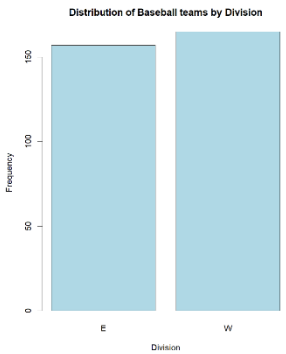
The data suggests substantial variability among players, with a few high-performing outliers driving the upper extremes.





4. Categorical Variable Analysis

The **Eastern (E)** and **Western (W)** divisions have nearly equal frequencies, with both divisions showing comparable counts around **160 teams each**. This balance in team distribution suggests an equitable split across divisions, potentially for competitive or geographical parity.



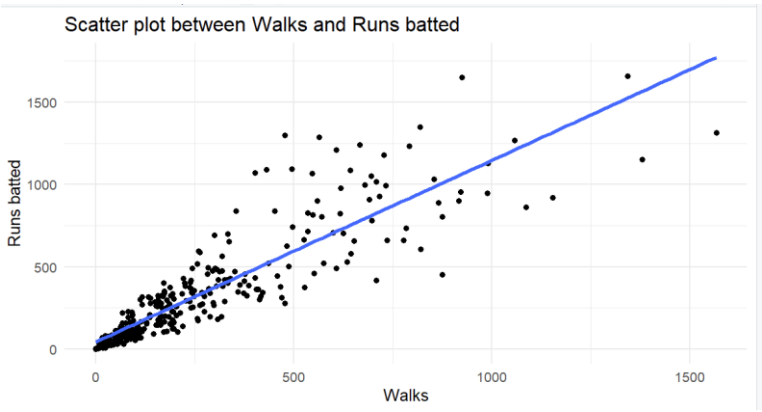
Multivariate Analysis Report

5. Correlation Analysis

Walks and Runs batted: A positive correlation (r = 0.732) indicates that as walks increase, prices tend to increase in that direction.

6. Scatter Plot Visualization

There is a clear positive relationship between **Walks** and **Runs batted**, indicating that as the number of walks increases, the number of runs batted also tends to rise. The correlation matrix confirms this trend with a **correlation of 0.732**, signifying a moderately strong linear association. Outliers are present, especially for higher values of **Walks** and **Runs**, which may warrant further investigation to understand exceptional cases.



7. Multiple Regression

Based on the correlation matrix and iterative approach, the final multiple regression model fitted is based on pages and subs values as follows:

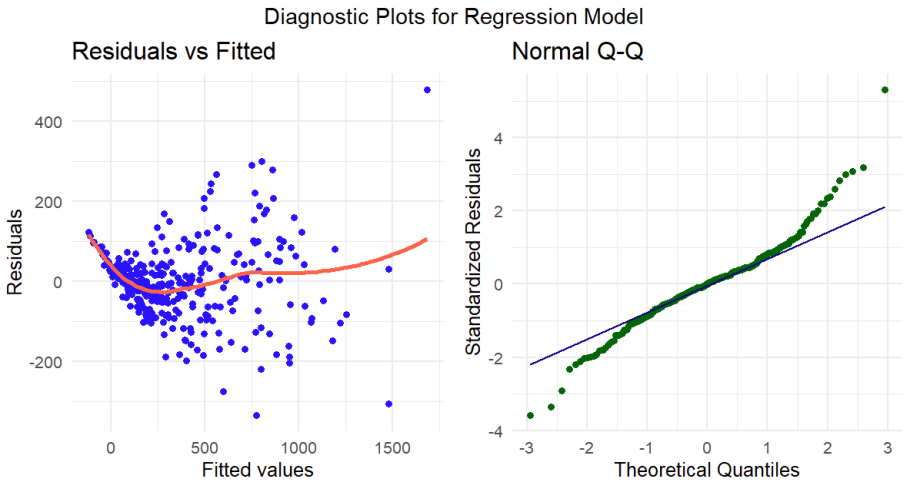
Runs = -155.26 + 27.18**(years)** +0.07**(walks)** + 0.33**(atbat86)**

The adjusted R2 is 0.920; the model explains ~92% of the variance in expenditure without a loss in predictive power. The RSE value turns out to be **94.9**.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -155.25641   17.06349  -9.099   <2e-16 ***
years         27.18547    2.00249  13.576   <2e-16 ***
walks          0.70836    0.03742  18.932   <2e-16 ***
atbat86        0.33428    0.03553   9.409   <2e-16 ***
---
```

8. Model Diagnostics

Residuals vs Fitted: The residual plot indicates a noticeable non-linear pattern, suggesting the model does not fully capture the relationship between predictors and the response variable. Residuals are more dispersed for higher fitted values, hinting at potential heteroscedasticity that may affect model accuracy.



Normal Q-Q Plot: The residuals deviate from the diagonal line, particularly at the extremes, indicating possible non-normality in the errors. Outliers are evident in both tails, which could influence model performance and suggest refinements such as transformations or robust regression methods.

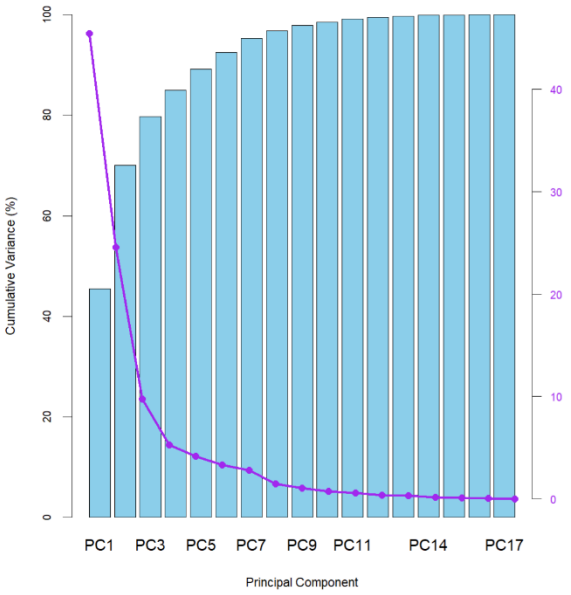
Advanced Analysis Report

9. Principal Component Analysis (PCA)

Explained Variance: The first two components explain ~70% of the variance, with PC1 alone accounting for ~45%. Adding PC3 increases the cumulative variance to ~80%, while subsequent components contribute diminishingly smaller portions.

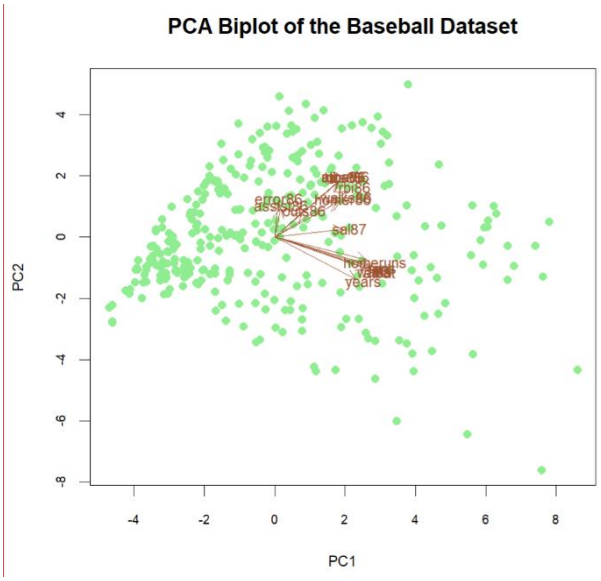
Component Selection: The scree plot highlights an "elbow" at PC2, suggesting 2 components are sufficient for a simplified analysis. Including PC3 could improve variance capture to ~80%.

Recommendation: Selecting 2 principal components strikes a balance between simplicity and variance retention while incorporating up to 3 components can justify a higher variance explanation.



10. PCA Interpretation

The biplot reveals that variables like homeruns, rbi86, walks86, and sal87 have strong contributions to PC1, as indicated by their longer vectors. Variables such as years and atbat86 also align closely with PC1, capturing key variability in player performance. PC2 shows moderate contributions from assist86 and outs86, differentiating some observations along this component. The dense clustering near the origin suggests most data points are similar in key performance metrics, while outliers spread along PC1 and PC2 highlight distinct patterns. This suggests PC1 dominates in explaining variability, with PC2 offering secondary insights into defensive statistics.



Conclusion

The **Baseball dataset** analysis provides key insights into player performance metrics. Univariate analysis reveals significant variability in runs, walks, and atbat86, with runs showing a right-skewed distribution and notable outliers. Categorical analysis indicates an even split between divisions E and W. Multivariate analysis uncovers strong positive correlations, particularly between walks, runs, and atbat86, reflecting their interdependence in player performance. The regression model explains much of the variability in runs but exhibits non-linear residual patterns, indicating scope for refinement. PCA effectively reduces dimensionality, with the first two components explaining ~70% of the variance, while three components capture ~80%. This analysis provides a strong foundation for identifying critical performance drivers, with room for further improvement in model fit and variable interactions.

APPENDIX

Credit Card Dataset

Summary Statistics (return)

VARIABLES	MEAN	SD	1ST QUARTILE	MEDIAN	3RD QUARTILE
ATBAT86	381	16	255	380	512
HITS86	101	1	64	96	137
HOMER86	10.8	0	4	8	16
RUNS86	50.9	0	30.2	48	69
RBI86	48	0	28	44	64.8
WALKS86	38.7	0	22	35	53
YEARS	7.44	1	4	6	11
ATBAT	2649	19	817	1928	3924
HITS	718	4	209	508	1059
HOMERUNS	69.5	0	14	37.5	90
RUNS	359	1	100	247	526
RBI	330	0	88.8	220	426
WALKS	260	0	67.2	170	339
OUTS86	289	0	109	212	325
ASSIST86	107	0	7	39.5	166
ERROR86	8.04	0	3	6	11
SAL87	536	67.5	190	425	750

Correlation Matrix (return)

	REPORTS	AGE	INCOME	SHARE	EXPENDITURE	DEPENDENTS	MONTHS	MAJORCARDS	ACTIVE
REPORTS	1	0.04408	0.011023	-0.15901	-0.13654	0.019731	0.048968	-0.0073	0.207755
AGE	0.044089	1	0.324653	-0.1157	0.014948	0.212146	0.436426	0.009777	0.18107
INCOME	0.011023	0.32465	1	-0.05443	0.281104	0.317601	0.130346	0.107138	0.18054
SHARE	-0.15901	-0.1157	-0.05443	1	0.838779	-0.08262	-0.05535	0.05147	-0.02347
EXPENDITURE	-0.13654	0.01495	0.281104	0.838779	1	0.052664	-0.02901	0.077514	0.054724
DEPENDENTS	0.019731	0.212146	0.317601	-0.08262	0.052664	1	0.046512	0.010285	0.107133
MONTHS	0.048968	0.436426	0.130346	-0.05535	-0.02901	0.046512	1	-0.04145	0.100028
MAJORCARDS	-0.0073	0.009777	0.107138	0.05147	0.077514	0.010285	-0.04145	1	0.119603
ACTIVE	0.207755	0.18107	0.18054	-0.02347	0.054724	0.107133	0.100028	0.119603	1

Diamonds Dataset

Summary Statistics (return)

VARIABLES	MEAN	SD	1ST QUARTILE	MEDIAN	3RD QUARTILE
CARAT	0.798	0.474	0.4	0.7	1.04
DEPTH	61.7	1.43	61	61.8	62.5
TABLE	57.5	2.23	56	57	59
PRICE	3933	3989	950	2401	5324
X	5.73	1.12	4.71	5.7	6.54
Y	5.73	1.14	4.72	5.71	6.54
Z	3.54	0.706	2.91	3.53	4.04

Correlation Matrix (return)

	CARAT	DEPTH	TABLE	PRICE	X	Y	Z
CARAT	1	0.028224	0.181618	0.921591	0.975094	0.951722	0.953387
DEPTH	0.028224	1	-0.29578	-0.01065	-0.02529	-0.02934	0.094924
TABLE	0.181618	-0.29578	1	0.127134	0.195344	0.18376	0.150929
PRICE	0.921591	-0.01065	0.127134	1	0.884435	0.865421	0.861249
X	0.975094	-0.02529	0.195344	0.884435	1	0.974701	0.970772
Y	0.951722	-0.02934	0.18376	0.865421	0.974701	1	0.952006
Z	0.953387	0.094924	0.150929	0.861249	0.970772	0.952006	1

Journals Dataset

Summary Statistics (return)

VARIABLES	MEAN	SD	1ST QUARTILE	MEDIAN	3RD QUARTILE
PRICE	418	386	134	282	541
PAGES	828	437	549	693	974
CHARPP	3233	819	2715	3010	3477
CITATIONS	647	1182	97.8	262	656
FOUNDINGYEAR	1967	25.7	1963	1973	1982
SUBS	197	205	52	122	268

Correlation Matrix (return)

	PRICE	PAGES	CHARPP	CITATIONS	FOUNDINGYEAR	SUBS
PRICE	1	0.493724	0.074579	0.028041	0.253416	-0.31197
PAGES	0.493724	1	-0.00899	0.537008	-0.15734	0.371406
CHARPP	0.074579	-0.00899	1	0.104458	0.03153	0.083193
CITATIONS	0.028041	0.537008	0.104458	1	-0.38304	0.584699
FOUNDINGYEAR	0.253416	-0.15734	0.03153	-0.38304	1	-0.40737
SUBS	-0.31197	0.371406	0.083193	0.584699	-0.40737	1

Baseball Dataset

Summary Statistics (return)

VARIABLES	MEAN	SD	1ST QUARTILE	MEDIAN	3RD QUARTILE
ATBAT86	381	16	255	380	512
HITS86	101	1	64	96	137
HOMER86	10.8	0	4	8	16
RUNS86	50.9	0	30.2	48	69
RBI86	48	0	28	44	64.8
WALKS86	38.7	0	22	35	53
YEARS	7.44	1	4	6	11
ATBAT	2649	19	817	1928	3924
HITS	718	4	209	508	1059
HOMERUNS	69.5	0	14	37.5	90
RUNS	359	1	100	247	526
RBI	330	0	88.8	220	426

WALKS	260	0	67.2	170	339
OUTS86	289	0	109	212	325
ASSIST86	107	0	7	39.5	166
ERROR86	8.04	0	3	6	11
SAL87	536	67.5	190	425	750

Correlation Matrix (return)

	ATBAT86	HITS86	HOMER86	RUNS86	RBI86	WALKS86	YEARS	ATBAT	HITS	HOMERUNS	RUNS	RBI	WALKS	OUTS86	ASSIST86	ERROR86	SAL87
ATBAT86	1	0.9679388 2	0.5921984 7	0.9130603 29	0.8205392	0.6698451	0.0473717 39	0.2355263 6	0.2527170 43	0.236659	0.2665339 8	0.2440531 3	0.1661231 5	0.3175498 1	0.3538239 8	0.3521171 86	0.3947709 45
HITS86	0.9679388 2	1	0.5621578 9	0.9221871 91	0.8110732 3	0.6412106	0.0447665 6	0.2275648 7	0.2558148 7	0.2026712	0.2617868 6	0.2320050 2	0.1518181 9	0.3106734 8	0.3204549	0.3100378 19	0.4386747 38
HOMER86	0.5921984 7	0.5621578 9	1	0.6509881 86	0.8551222	0.4810143	0.1163183 33	0.2218821 02	0.2206266 422	0.4932269	0.2623606 1	0.3519790 1	0.2331536 7	0.2829229 3	-0.106328606	0.0393176 54	0.3430280 78
RUNS86	0.9130603 3	0.9221871 9	0.6509881 9	1	0.7982056 67	0.7322128 48	0.0045412 71	0.1864973 88	0.2048297 12	0.2279132	0.2505560 9	0.2059759 3	0.1821682	0.2793472 2	0.2205673 54	0.2404750 25	0.4198585 59
RBI86	0.8205392 3	0.8110732 3	0.8551222	0.7982056 66	1	0.6159971	0.1461681 28	0.2946883 74	0.3082010 64	0.4417712	0.3232846 4	0.3931837 2	0.2509136	0.3431864 1	0.1065910 21	0.1936973 1	0.4494570 88
WALKS86	0.6698451	0.6412106	0.4810143	0.7322128 48	0.6159971	1	0.1364749 71	0.2771747 91	0.2806710 54	0.3324732	0.3384779 8	0.3086313 9	0.4245070 9	0.2995146 9	0.1496561 13	0.1293820 59	0.4438672 6
YEARS	0.0473717 4	0.0447665 6	0.1163183 33	0.0045412 71	0.1461681	0.1364749 71	1	0.9202893 61	0.9030631 057	0.7268721	0.8828769 1	0.8688121 4	0.8385330 4	-0.00468406	-0.080638434	-0.162140389	0.4006569 94
ATBAT	0.2355263 6	0.2275648 7	0.2218821	0.1864973 88	0.2946883	0.2771747	0.9202893 61	1	0.9950634 82	0.7988364	0.9833453 9	0.9492187 8	0.9065006 5	0.0622828 5	0.0020378 39	-0.066921843	0.5261353 1
HITS	0.2527170 4	0.2558148 7	0.2206266 4	0.2048297 12	0.3082010	0.2806710	0.9030631 057	0.9950634 82	1	0.7833064	0.9846088 2	0.9451410 2	0.8909540 4	0.0765467 4	-0.002522916	-0.062756073	0.5489095 59
HOMERUNS	0.2366589 9	0.2027119 6	0.4932269 2	0.2279131 99	0.4417712	0.3324732	0.7268721	0.7988364	0.7833064	1	0.8202427 5	0.9294837 3	0.7999828 6	0.1127244 2	-0.158510 87	-0.138114787	0.5249305 6
RUNS	0.2665339 8	0.2617868 6	0.2623606 1	0.2505560 94	0.3232846	0.3384779	0.8828769 1	0.9833453 9	0.9846088 6	0.8202428	1	0.9437690 2	0.9278069 2	0.0641795 9	-0.022978 15	-0.084395119	0.5626777 11
RBI	0.2440531 3	0.2320050 2	0.3519790 1	0.2059759 31	0.3931837	0.3086313	0.8688121 4	0.9492187 8	0.9451410 2	0.9294837	0.9437690 2	1	0.8847258 2	0.1100982 9	-0.079387341	-0.100989953	0.5669656 86
WALKS	0.1661231 5	0.1518181 9	0.2331536 7	0.1821682	0.2509136	0.4245070	0.8385330 4	0.9065006 5	0.8909540 4	0.7999829	0.9278069 2	0.8847258 2	1	0.0586378 9	-0.039129763	-0.118474728	0.4898220 36
OUTS86	0.3175498 1	0.3106734 8	0.2829229 3	0.2793472 21	0.3431864	0.2995146	-0.00468406	0.0622828 5	0.0765467 4	0.1127244	0.0641795 9	0.1100982 9	0.0586378 9	1	-0.025024002	0.1099715 43	0.3004803 56
ASSIST86	0.3538239 8	0.3204549	-0.106328 61	0.2205673 54	0.1065910	0.1496561	-0.080638434	0.0020378 39	-0.002522916	-0.158 5109	-0.022978 15	-0.079387341	-0.039129763	-0.025024	1	0.7063620 22	0.0254361 36
ERROR86	0.3521171 19	0.3100378 2	0.0393176 5	0.2404750 25	0.1936973	0.1293820	-0.162140389	-0.066921843	-0.062756073	-0.138 1148	-0.084395 12	-0.100989953	-0.118474743	0.1099715 4	0.7063620 22	1	-0.005400702
SAL87	0.3947709 4	0.4386747 4	0.3430280 8	0.4198585 59	0.4494570	0.4438672	0.4006569 94	0.5261353 1	0.5489095 59	0.5249306	0.5626777 1	0.5669656 9	0.4898220 4	0.3004803 6	0.0254361 36	-0.005400702	1