TITANIC

Univariate Analysis

Titanic dataset has 8 features and there is total 891 observations. Sex, Ticket, cabin, Embarked are categorical columns where Age, Fare, SibSp and Parch are numerical columns. SibSp is total number of siblings and spouse, whereas Parch is total number of parents and child.

Summary of numerical columns

	Age	Fare	Parch	SibSp
Min	0.42	0.00	0.0000	0.0000
1st Qu	20.12	8.05	0.0000	0.0000
Median	28.00	15.74	0.0000	0.0000
Mean	29.70	34.69	0.4314	0.5126
3rd Qu	38.00	33.38	1.0000	1.0000
Max	80.00	512.33	6.0000	5.0000

- Histogram of Age shows nearly normal distribution with most number of people in age between 20 and 30.
- Fare seems to follow exponential distribution, as amount of fare increases, number of frequencies decreases.
- Columns Sibsp and Parch seems to follow exponential distribution, where 1 are most frequently observed values for both variable.
- Age variable have some outliers where age is more than 60. There are more number of outliers for fare column where values greater than 100 are considered as outliers. For Parch and SibSp column, values greater than 3 are considered as an outlier.

Categorical variable Analysis

- More numbers of people are not able to survive. More number of people are travelling from class 3 followed by class 1 and class 2.
- There are more number of males than females in this ship.
- 0 and 1 are the most frequent values for SibSp and Parch variables.

Multivariate Analysis

Correlation Analysis

Variable 1	Variable 2	Pearson Correlation
Age	Fare	0.09606669
Age	SibSp	-0.3082468
Age	Parch	-0.1891193
Fare	SibSp	0.1383288
Fare	Parch	0.2051189
SibSp	Parch	0.3838199

- ➔ Age has very slightly positive correlation with fare that means fare doesn't depend much on age linearly. And as age increases, SibSp and Parch decreases.
- ➔ Fare has positive correlation with SibSp and Parch and SibSp has positive correlation with Parch.

Numeric – Categorical Variables

- ➔ Female has paid more average fare than male. Average fare are higher for embark C followed by S and Q. Embark C has more number of average SibSp compared to Embark S and Q. Average SibSp and Parch for females are more than male that means more number of males likes to travel alone.

Categorical – Categorical Variables

Sex / Embarked	C	Q	S
Female	61	12	186
Male	69	16	368

Multiple Linear Regression

- ➔ Categorical variable Sex is one-hot encoded into new variable female that means if Sex is female, then value of female will be 1 otherwise 0. Similarly, from variable Embarked, 2 new columns are created: embarked_C and embarked_q. If Embarked has value C, then embarked_C will be 1. If Embarked has value Q, then embarked_q will be 1. If both embarked_C and embarked_q are 0, that means Embarked has value S.
- ➔ The model predicts Fare using Age, SibSp, Parch, female, and two embarkation indicators. All variables except embarked_q are significant (p-values < 0.05). Each additional sibling or spouse increases the fare by 7.47, and being female increases it by 13.15. The Adjusted R-squared value (0.1699) means that the model explains about 17% of the variance in Fare.

Model Summary

Residuals	Min	1Q	Median	3Q	Max
	-74.18	-18.51	-7.26	1.86	454.92

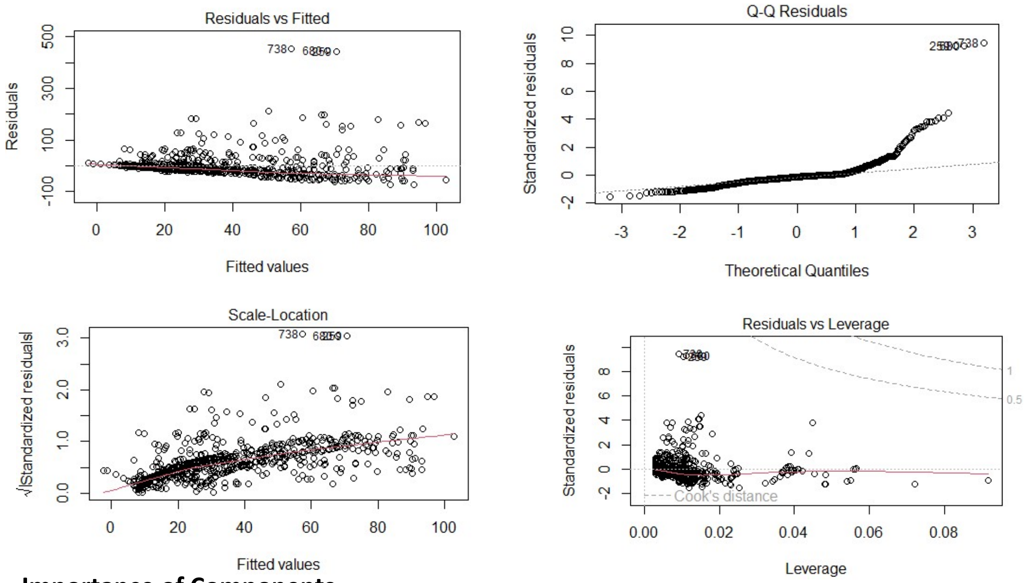
Residual Standard error: 48.21 on 707 degrees of freedom, Multiple E-squared: 0.1769, Adjusted R-squared: 0.1699 and F-statistic: 25.32 on 6 and 707 DF, p-value < 2.2e-16

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-2.8874	5.0225	-0.575	0.565550
Age	0.6065	0.1313	4.619	4.57e-06
SibSp	7.4730	2.1824	3.424	0.000652
Parch	9.8476	2.3596	4.173	3.38e-05
female	13.1507	3.8974	3.374	0.000781
embarked_C	39.0713	4.7322	8.256	7.36e-16
embarked_q	-10.9932	9.3606	-1.174	0.240625

Model Diagnostics

-> The Residuals vs. Fitted plot shows heteroscedasticity as residuals spread unevenly with non-constant variance. The Q-Q plot shows deviations from normality, especially in the tails. The Scale-Location plot also confirms heteroscedasticity.

-> Residuals vs. Leverage plot shows a few influential points suggesting outliers affecting the model. Overall, while some predictors are significant, the assumptions of linear regression are violated.



PCA - Importance of Components

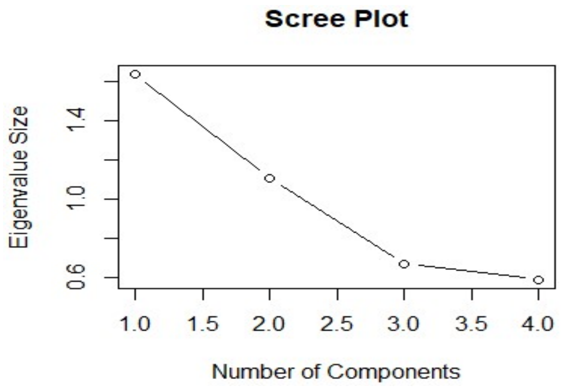
	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Standard deviation	1.2793554	1.0522248	0.8181718	0.7659423
Proportion of Variance	0.4091876	0.2767942	0.1673513	0.1466669
Cumulative Proportion	0.4091876	0.6859818	0.8533331	1.0000000

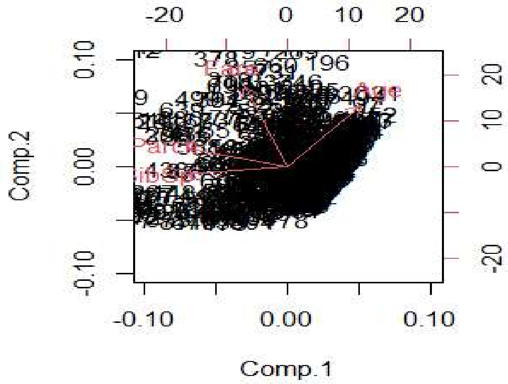
Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Age	0.439	0.596	0.561	0.370
SibSp	-0.625			0.775
Parch	-0.591	0.177	0.606	-0.503
Fare	-0.260	0.780	-0.562	

PCA was conducted on four numerical variables (Age, SibSp, Parch, Fare). The scree plot suggests selecting two principal components as their eigenvalues are above 1 which explains 68.6% of the total variance (40.9% by PC1 and 27.7% by PC2).

Biplot

The biplot highlights the relationship between variables and observations in terms of the first two components. Fare and Age have strong contributions to PC2 and PC1, respectively. Variables SibSp and Parch load negatively on PC1, reflecting similarity in their behavior. Observations with higher SibSp and Parch values cluster on the negative side of PC1, while those with high Fare group along PC2. The biplot shows clear groupings influenced by these relationships, indicating distinct patterns in passenger demographics and ticket fares.





AIRQUALITY

Univariate Analysis

AirQuality has total 153 observations across 6 variables. Type of all variables are numerical but day of month and month number can be considered as categorical where day of month could be any value between 1 to 31 and month number could be anything between 5 and 9.

Summary of numerical columns

	Ozone	Solar.R	Wind	Temp
Min	1.00	7.0	2.30	57.00
1st Qu	18.00	113.5	7.40	71.00
Median	31.0	207.0	9.70	79.00
Mean	42.1	184.8	9.94	77.79
3rd Qu	62.0	255.5	11.50	84.50
Max	168.0	334.0	20.70	97.00

- Lower value has high frequency in the Ozone variable whereas higher value has lower frequencies and it seems to follow exponential distribution. For Solar.R, frequencies increase as value of variable increases
- As temperature increases, number of observations increase and then decreases with increase in temperature. Ozone column has 2 outliers and wind has 3 outliers

Categorical variable Analysis

- For month column, all observations are from month 5 to 9. And month 6 and 8 has lowest number of observations which means there were many missing values for those and those are removed.

Multivariate Analysis - Correlation Analysis

➔ As ozone has positive correlation with Solar.R and Temp means that as ozone level increases, solar radiation and temperature increases. And ozone has negative correlation with wind means as ozone level increases, wind speed decreases.

- Solar radiation has slightly negative (close to zero) correlation with wind, which means that solar radiation doesn’t have much influence on wind. And it has positive correlation with temperature means as solar radiation increases, temperature increases.
- Wind has negative correlation with temperature, that means as wind speed increases, temperature decreases.

Variable 1	Variable 2	Pearson correlation
Ozone	Solar.R	0.3483417
Ozone	Wind	-0.6124966
Ozone	Temp	0.6985414
Solar.R	Wind	-0.1271835
Solar.R	Temp	0.2940876
Wind	Temp	-0.4971897

Numeric – Categorical Variables

- Average Ozone level starts to increase between May and August and then it starts to fall down in September. From June to July, rise in ozone level was sharpest.
- Average Solar radiation increases from May to July and then it falls for August and September. Rise from June to July was sharpest and decline from July to August is also sharpest.
- Average Wind speed increase from May to June, then it falls in July and again rises in August and September. High average wind speed is observed in June month. There is sharp decline in average wind speed from June to July.
- From May to July, average temperature rises sharply, and there is slight decline in temperature in August followed by sharp decline in September.

Multiple Regression

- A multiple regression model is fit to predict Temp using Ozone, Solar.R, Wind, and one-hot encoded month variables (May, June, July, August). The summary shows the relationship between predictors and the dependent variable. The coefficient for Wind shows its negative influence on temperature, while the month variables highlight seasonal variations.

Model Summary

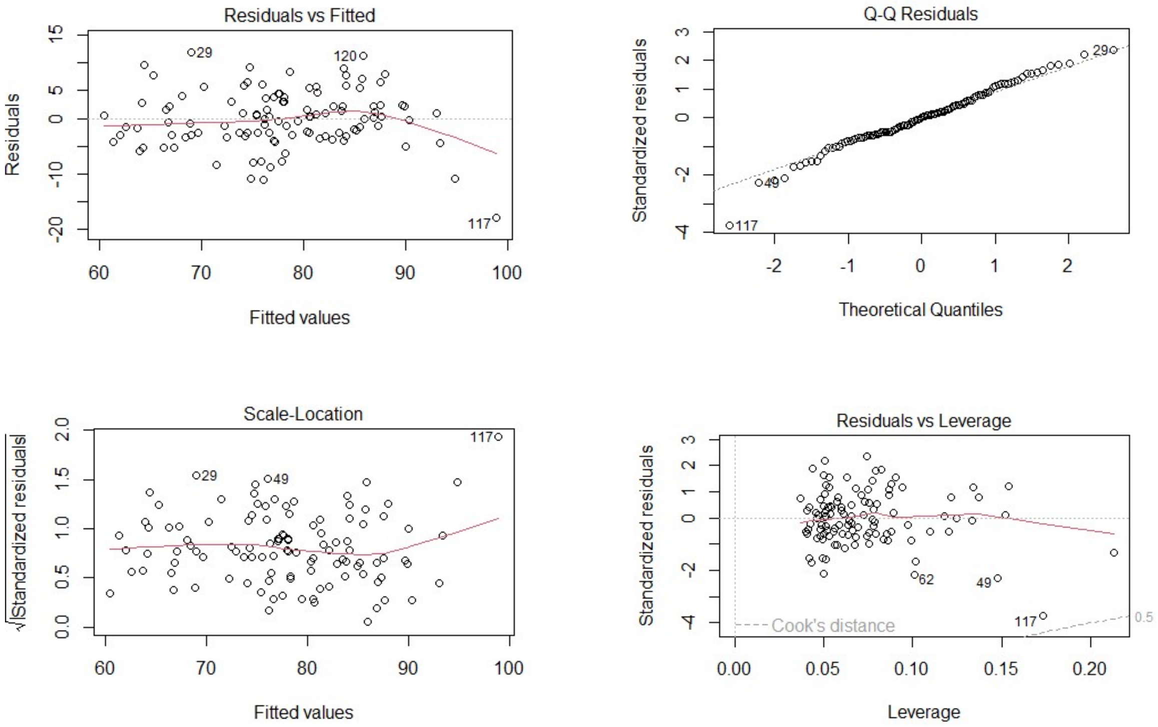
Residuals	Min	1Q	Median	3Q	Max
	-17.9220	-3.0386	0.0148	3.0856	12.0292

Residual standard error: 5.268 on 103 degrees of freedom, Multipe R-squared: 0.7139, Adjusted Q-squared: 0.6944 and F-statistic: 36.71 on 7 and 103 DF, p-value: < 2.2e-1

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	73.605129	2.527716	29.119	<2e-16
Ozone	0.121180	0.022020	5.503	2.74e-07
Solar.R	0.011901	0.006044	1.969	0.0517

Wind	0.250226	0.183341	-1.365	0.1753
May	9.358031	1.473927	-6.249	5.93e-09
June	1.903854	2.043538	0.932	0.3537
July	2.673023	1.504135	1.777	0.0785
August	2.977115	1.562692	1.905	0.0596

Model Diagnostics: Residual diagnostic plots shows reasonable model assumptions. The Residuals vs. Fitted plot shows no clear pattern, which means it is homoscedastic. The Q-Q plot confirms the normality of residuals except for slight deviations for extreme points. But leverage and Cook's distance plots suggest some influential observations that requires closer inspection.



Principal Component Analysis (PCA)

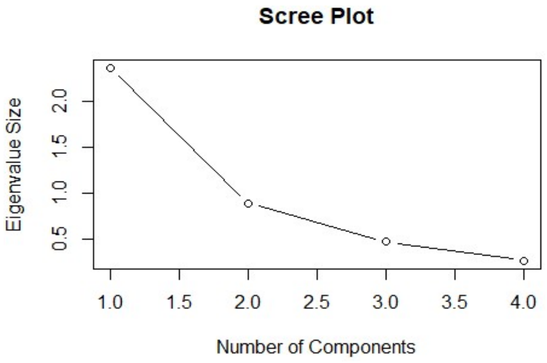
Explained Variance: PCA performed on numerical variables (Ozone, Solar.R, Wind, Temp) shows two components with eigenvalues close to 1, explaining about 80% of the total variance. Thus, retaining two components balances dimensionality reduction and information retention.

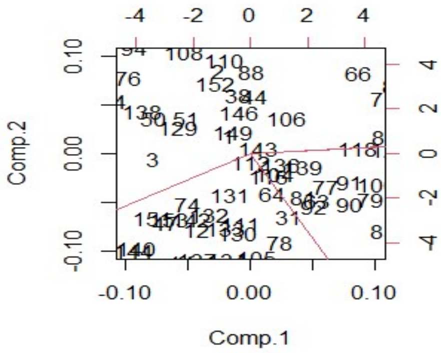
PCA - Importance of components

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Standard deviation	1.5361961	0.9458733	0.6897463	0.5193026
Proportion of Variance	0.5899747	0.2236691	0.1189375	0.0674188
Cumulative Proportion	0.5899747	0.8136437	0.9325812	1.0000000

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Ozone	0.589		0.115	0.797
Solar.R	0.317	-0.898	-0.278	-0.123
Wind	-0.497	-0.430	0.691	0.302
Temp	0.553		0.658	-0.508

PCA Interpretation: The biplot shows strong contributions of Wind and Ozone to PC1 and PC2. Variables cluster along distinct axes, showing correlations. Observations with higher Solar.R and Temp and are grouped distinctly which shows seasonal or environmental patterns.





PENGUINS

Univariate Analysis

Penguins has 345 observations across 7 total variables with mixed types (Numeric and categorical). Species, island and sex is categorical variable whereas bill_length_mm, bill_depth_mm, flipper_length_mm and body_mass_g are numerical variables.

Summary of numerical columns

	Bill_length_mm	Bill_depth_mm	flipper_length_mm	body_mass_g
Min	32.10	13.10	172.0	2700
1st Qu	39.23	15.60	190.0	3550
Median	44.45	17.30	197.0	4050
Mean	43.92	17.15	200.9	4202
3rd Qu	48.50	18.70	213.0	4750
Max	59.60	21.50	231.0	6300

- bill_length_mm and bill_depth_mm has lower frequencies on extreme sides and frequencies are similar and higher for mid-range values which are not at the extremes.
- For body_mass_g histogram, frequencies reduces as values are more for body mass in gram
- Based on the boxplot, there are not any outliers in penguin dataset for all these 4 numerical columns

Categorical variable Analysis

- There are 3 categories for penguin species: Adelie, Chinstrap and Gentoo. Adelie are more common penguins followed by Gentoo. Whereas Chinstrap are least in numbers.

- Penguins are seen in 3 different islands. Biscoe has most number of penguins, dream has less number of penguin compared to Dream and Torgersen has least number of penguins among all 3 islands.
- Based on sex, there aren’t much difference between number of female and number of male penguins.

Multivariate Analysis

Correlation Analysis

Variable 1	Variable 2	Pearson correlation
bill_length_mm	bill_depth_mm	-0.2350529
bill_length_mm	flipper_length_mm	0.6561813
bill_length_mm	body_mass_g	0.5951098
bill_depth_mm	flipper_length_mm	-0.5838512
bill_depth_mm	body_mass_g	-0.4719156
flipper_length_mm	body_mass_g	0.8712018

➔ As bill_length_mm increases, flipper_length_mm and body_mass_g increases. And as bill_length_mm increases bill_depth_mm decreases.

➔ As bill_depth_mm increases both flipper_length_mm and body_mass_g decreases.

➔ flipper_length_mm has very high positive correlation body_mass_g.

Numeric – Categorical Variables

➔ Chinstrap species has highest average bill depth followed by Adelie and Gentoo species.

➔ Average Body mass for species Gentoo is higher than other 2 species and there is almost similar average body mass for both Adelie and Chinstrap.

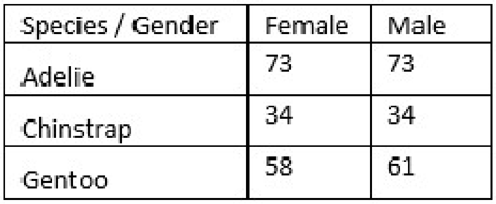
➔ Gentoo has higher average flipper length compared to other 2 species. Average bill length for Chinstrap is the highest among 3 species.

➔ There is not much difference between average bill length, average bill depth and average flipper length based on sex. Whereas male penguin has higher average body mass than female penguin.

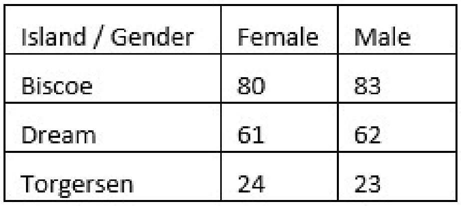
➔ Based on island as well, there is not much difference in average bill length, average bill depth and average flipper length between species living on 3 islands. But average body mass of penguin living on Biscoe island are significantly higher compared to penguin living on Torgerson and Dream island.

Categorical – Categorical Variables

Species / Island	Biscoe	Dream	Torgersen
Adelie	44	56	51
Chinstrap	0	68	0
Gentoo	123	0	0



Species / Gender	Female	Male
Adelie	73	73
Chinstrap	34	34
Gentoo	58	61



Island / Gender	Female	Male
Biscoe	80	83
Dream	61	62
Torgersen	24	23

Multiple Linear Regression

The multiple regression analysis provides valuable insights into the factors influencing penguin body mass. Bill dimensions (length and depth) and flipper length emerge as significant predictors, underscoring their strong correlation with body mass. Species play a critical role, with Gentoo penguins serving as the heaviest baseline, while Adelie and Chinstrap species show significant reductions in body mass. Female penguins are consistently lighter than males, reflecting sexual dimorphism in the species. However, island location has limited influence on body mass, as neither Torgersen nor Dream islands significantly affect the results. The model performs robustly, explaining nearly 87% of the variance in body mass.

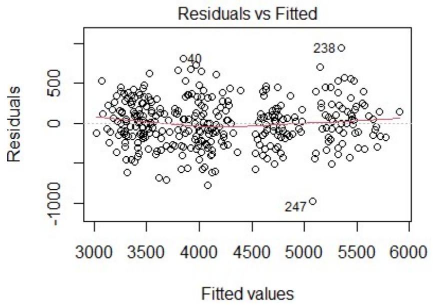
Model Summary

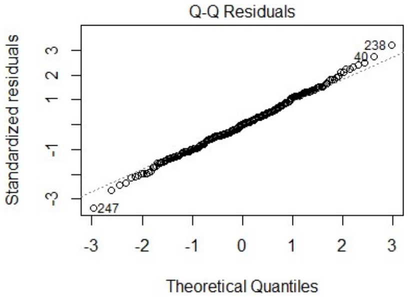
Residuals	Min	1Q	Median	3Q	Max
	-980.2	-176.3	-1.3	181.8	948.1

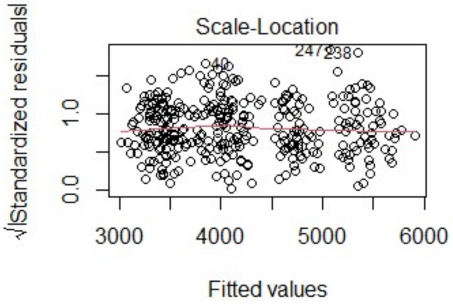
Coefficients

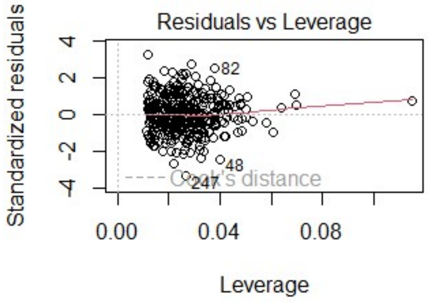
Variables	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	-850.157	652.412	-1.303	0.193442
bill_length_mm	24.324	7.169	3.393	0.000776
bill_depth_mm	79.317	19.890	3.988	8.20e-05
flipper_length_mm	17.193	2.972	5.786	1.66e-08
Adelie	-930.321	139.436	-6.672	1.05e-10
Chinstrap	-1249.308	131.336	-9.512	<2e-16
Torgersen	-58.124	61.366	-0.947	0.344239
Dream	-22.198	59.982	-0.370	0.711556
FEMALE	-321.999	46.770	-6.885	2.89e-11

Residual standard error: 296.4 on 333 degrees of freedom, Multiple R-squared: 0.8666, Adjusted R-squared: 0.8634 and F-statistic: 270.4 on 8 and 333 DF, p-value: <2.2e-16









Model Diagnostic

Diagnostic plots show residuals are approximately normally distributed, though minor heteroscedasticity is observed, as indicated by a slight spread in residual patterns. Overall, the model fits well, accurately capturing the relationships between predictors and body mass, but the slight deviation from assumptions.

Principal Component Analysis (PCA)

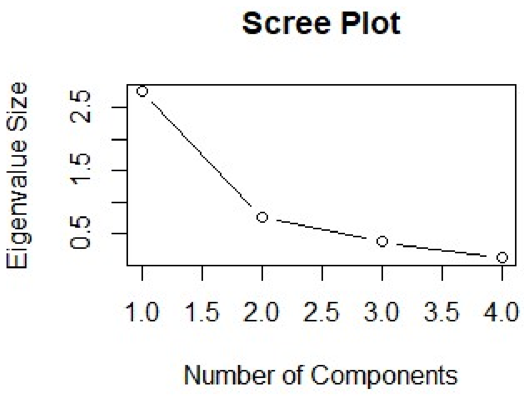
The PCA highlights the dominant patterns in the penguin dataset while reducing its complexity. Two principal components explain the majority of the variance, with the first capturing size-related features like body mass, bill length, and flipper length, and the second emphasizing bill depth. The scree plot confirms the appropriateness of focusing on the first two components, as additional components contribute marginally to the variance.

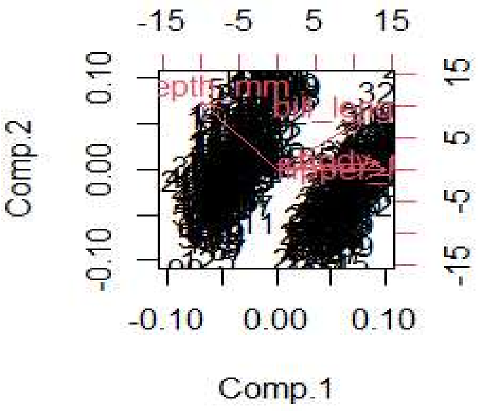
PCA - Importance of components

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Standard deviation	1.6594442	0.8789293	0.60434750	0.32938157
Proportion of Variance	0.6884388	0.1931292	0.09130898	0.02712305
Cumulative Proportion	0.6884388	0.8815680	0.97287695	1.00000000

Loadings	Comp. 1	Comp. 2	Comp. 3	Comp. 4
bill_length_mm	0.455	0.597	0.644	0.146
bill_depth_mm	-0.400	0.798	-0.418	-0.168
flipper_length_mm	0.576		-0.232	-0.784
body_mass_g	0.548		-0.597	0.580

The biplot reveals distinct relationships among variables: heavier and larger penguins cluster together, with flipper length strongly aligned with body mass, while bill depth diverges somewhat in its influence. This dimensionality reduction approach also hints at species separations, as clusters in the PCA space may reflect differences in body size and shape among penguins. PCA offers a clear visualization of interrelationships between traits, simplifying the understanding of complex morphological variations within the dataset.





TIPS

Univariate Analysis

Tips has 245 observations across 7 total variables with mixed types numerical and categorical. Total_bill and tip are numerical columns whereas size, sex, smoker, day and time are categorical variables.

Summary of numerical columns

	total_bill	Tip	size
Min	3.07	1.000	1.00
1st Qu	13.35	2.000	2.00
Median	17.80	2.900	2.00
Mean	19.79	2.998	2.57
3rd Qu	24.13	3.562	3.00
Max	50.81	10.000	6.00

- Variable tip looks like it follows exponential distribution, where initial lower values have high frequencies and as values grow up, its frequencies decreases.
- 2 is common size in this dataset. total_bill has lower frequencies on extreme sides and frequencies are similar and higher for mid-range values which are not at the extremes.
- There are 7 outliers in total_bill column where values are more than 40 and 6 outliers in tip column when values are more than 6. Size columns have 2 outliers for which values are more than 4.

Categorical variable Analysis

- There are more number of males compared to female which shows that male pays the bill more times than female.
- Number of people who smoke are less than number of people who don’t smoke.
- Only 4 days are there in the dataset: Thursday, Friday, Saturday and Sunday. Most number of bills are paid on Saturday followed by Sunday, Thursday and Friday respectively.
- People who paid for dinner are more than people who paid for lunch.
- Size varies from 1 to 6 and maximum number of observations are with size 2.

Multivariate Analysis

Correlation Analysis

Variable 1	Variable 2	Pearson correlation
total_bill	tip	0.6757341
total_bill	size	0.5983151
tip	size	0.4892988

➔ All three variables have positive correlation with each other that means all three variables total_bill, tip and size increases together.

Multiple Linear Regression

- The categorical variable sex is one hot encoded into Female. Similarly, smoker is one hot encoded where 1 means person is smoker and 0 means person is not smoker. Column time is one hot encoded into dinner, where for dinner value 1 means it bill is for dinner time and 0 means bill is for lunch time. And day is one hot encoded into 3 variables sunday, saturday and thursday as it has got 4 values.
- The multiple regression analysis predicts tip based on total_bill, size, and categorical variables like Female, smoker, dinner, and specific days of the week.
- The model explains 47% of the variance in tips (Adjusted $R^2 = 0.452$) with a statistically significant overall F-test ($p < 2.2e-16$). Among predictors, total_bill is the most significant, with a positive relationship (Estimate = 0.094, $p < 0.001$), indicating that tips increase proportionally with higher bills.
- Size is marginally significant (Estimate = 0.176, $p = 0.0505$), suggesting that larger group sizes slightly raise tips. Other variables like Female, smoker, and meal timing (dinner) are not statistically significant. day-of-week effects (e.g., sunday, saturday, and thursday) are negligible.
- Residual standard error (1.024) indicates moderate variability unexplained by the model. Overall, the model demonstrates the importance of bill size and group size while showing minimal effects for other predictors.

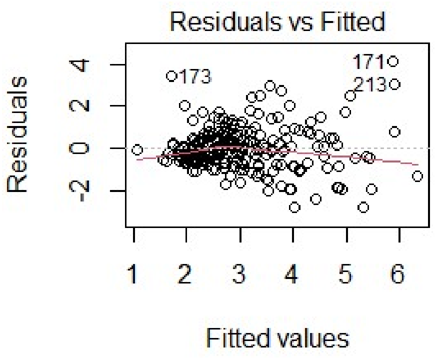
Model Summary

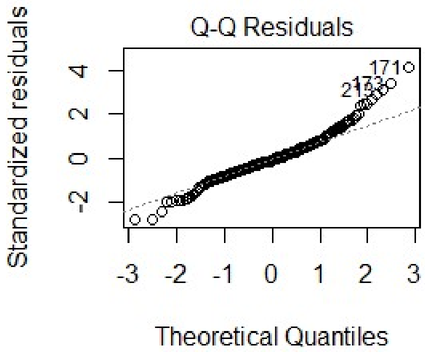
Residuals	Min	1Q	Median	3Q	Max
	-2.8475	-0.5729	-0.1026	0.4756	4.1076

Coefficients

Variables	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept)	0.839505	0.424024	1.980	0.0489
total_bill	0.094487	0.009601	9.841	<2e-16
size	0.175992	0.089528	1.966	0.0505
Female	0.032441	0.141612	0.229	0.8190
smoker	-0.086408	0.146587	-0.589	0.5561
dinner	-0.068129	0.444617	-0.153	0.8783
sunday	-0.025481	0.321298	-0.079	0.9369
saturday	-0.121458	0.309742	-0.392	0.6953
thursday	-0.162259	0.393405	-0.412	0.6804

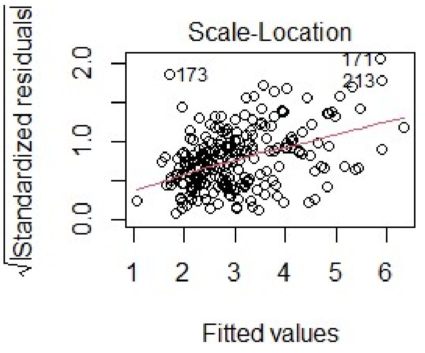
Residual standard error: 1.024 on 235 degrees of freedom, Multiple R-squared: 0.4701, Adjusted R-squared: 0.452 and F-statistic: 26.06 on 8 and 235 DF, p-value: <2.2e-16

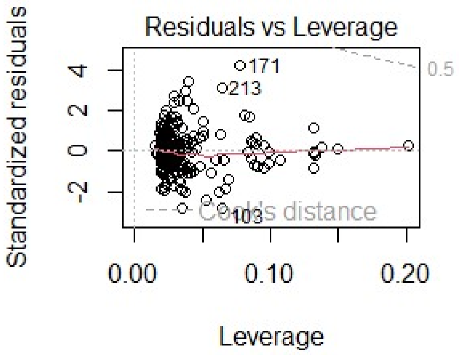




Model Diagnostics

Residual diagnostic plots evaluate the model fit. The "Residuals vs. Fitted" plot shows a scattered distribution around zero, indicating no clear pattern and supporting linearity. The Q-Q plot suggests residuals are roughly normally distributed, with minor deviations at the tails. The "Scale-Location" plot shows slight heteroscedasticity, where variance increases with fitted values. However, this is not severe enough to undermine the model's validity. The "Residuals vs. Leverage" plot identifies a few points (e.g., 1710 and 2130) with high leverage, but Cook’s distance indicates they do not unduly influence the model. Overall, the diagnostic plots confirm that the model meets key assumptions, although slight improvements could enhance fit.





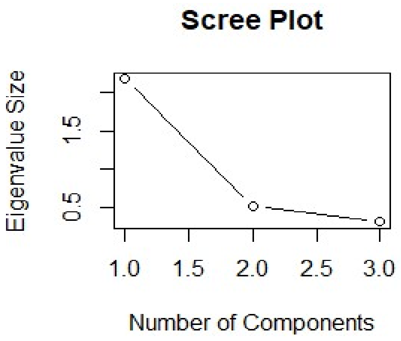
PCA - Importance of components:

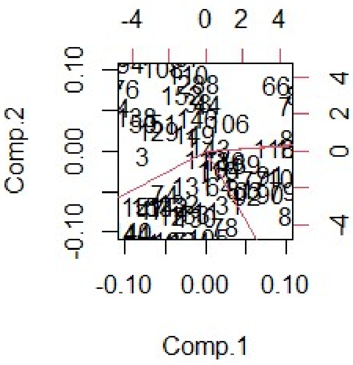
	Comp. 1	Comp. 2	Comp. 3
Standard deviation	1.4760859	0.7205046	0.5495848
Proportion of Variance	0.7262766	0.1730423	0.1006811
Cumulative Proportion	0.726766	0.893189	1.000000

PCA on numerical variables (total_bill, tip, size) reduces dimensionality effectively. The first two components explain 89.3% of the variance(72.6% by PC1, 17.3% by PC2), as shown in the scree plot. Retaining these two components balances simplicity and information preservation. Loadings reveal that PC1 primarily represents overall size, with strong contributions from total_bill (0.608), tip(0.576), and size(0.547). PC2 contrasts size(-0.791) against tip(0.593), indicating differences in group sizes and tipping behavior.

Loadings:	Comp. 1	Comp. 2	Comp. 3
total_bill	0.608	0.151	0.780
tip	0.576	0.593	-0.563
size	0.547	-0.791	-0.274

The biplot visualizes these relationships, showing clustering of observations and highlighting outliers. The PCA simplifies the dataset while retaining key patterns, such as the association between higher bills and tips and variability in tipping for larger groups.


Scree Plot
Eigenvalue Size
Number of Components


Comp.2
Comp.1