Statistical Structures in Data
Prof. Subhajit Dutta

Mohsin Parmar
24BM6JP33

SLEEP IN MAMMALS

1.Overview of Data Set

	species	body_wt	brain_wt	non_dreaming	dreaming	total_sleep	life_span	gestation
1	Africanelephant	6654.000	5712.0	8.672917	1.972	3.3	38.60000	645
2	Africangiantpouchedrat	1.000	6.6	6.300000	2.000	8.3	4.50000	42
3	ArcticFox	3.385	44.5	8.672917	1.972	12.5	14.00000	60
4	Arcticgroundsquirrel	0.920	5.7	8.672917	1.972	16.5	19.87759	25
5	Asianelephant	2547.000	4603.0	2.100000	1.800	3.9	69.00000	624
6	Baboon	10.550	179.5	9.100000	0.700	9.8	27.00000	180

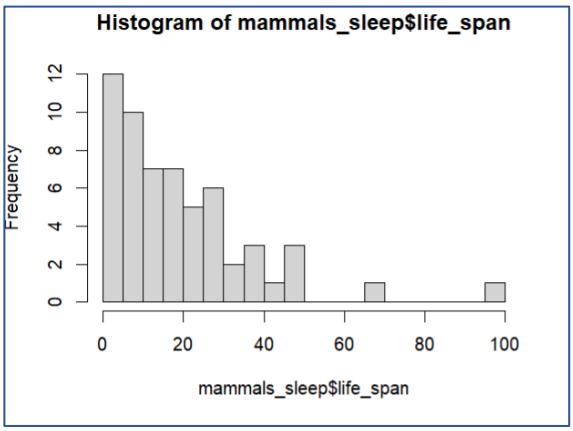
Number of observations: **62** Number of Variables : **11**

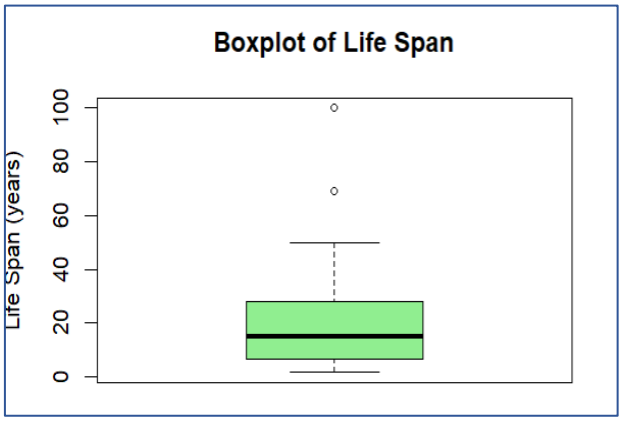
2. Summary Satistics for life span

Mean: 19.87759 **Median:** 15.1 **Standard deviation:** 18.20626 **Minimum :** 2 **Maximum :** 100

3. Distribution Visualization for Lifespan

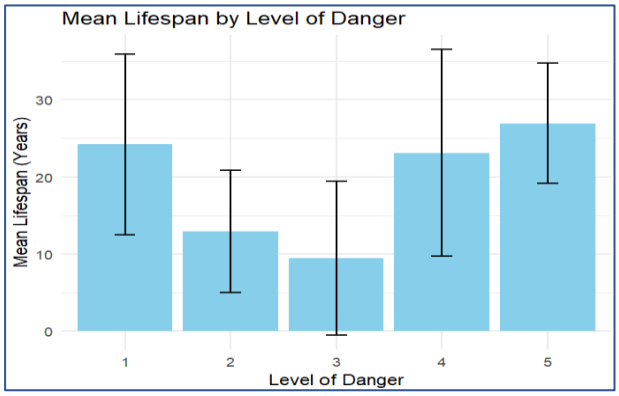
The histograms and boxplots illustrate the distribution of Lifespan. Both the histogram and the barplot show 2 potential outliers on the higher side. From the Histogram the distribution seems to be **exponential**(decay).





4. Categorical Variable Analysis :

The distribution of categorical variables is as follows: For 'danger', there seems to be a higher count for certain categories. This suggests possible relationships between categorical features and other variables like sleep patterns.
Summary: Animals having a danger of level 3 have lesser mean life span.



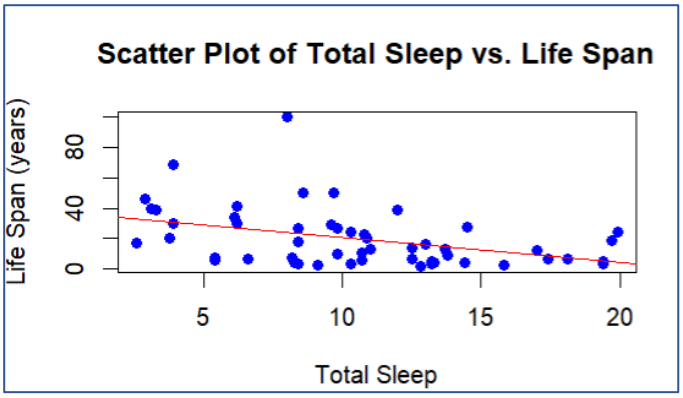
Mohsin Parmar
24BM6JP33

5. Correlation Analysis:

Pearson correlation coefficient between life span and total_sleep: **-0.4102024**
Summary: The Pearson correlation suggests a **negative relationship** between life span and total sleep.

6. Scatterplot Visualization:

The scatterplot shows a negative relationship between life span and total sleep, consistent with the Pearson correlation coefficient of -0.4102. The trendline highlights this inverse relationship.



7.Multiple Linear Regression.

model <- lm(life_span ~ body_wt + gestation, data = mammals_sleep)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.575040465	2.999333396	2.525575	1.463723e-02
body_wt	-0.003230753	0.002749517	-1.175026	2.453369e-01
gestation	0.089668809	0.017644245	5.082043	5.165425e-06

	R_Squared	Adjusted_R_Squared
1	0.394126	0.3708231

	Min	Q1	Median	Mean	Q3	Max	Std_Dev	IQR
25%	-22.62163	-7.371896	-2.424136	-1.113377e-17	5.890604	68.68369	14.31189	13.2625

Key Interpretations:

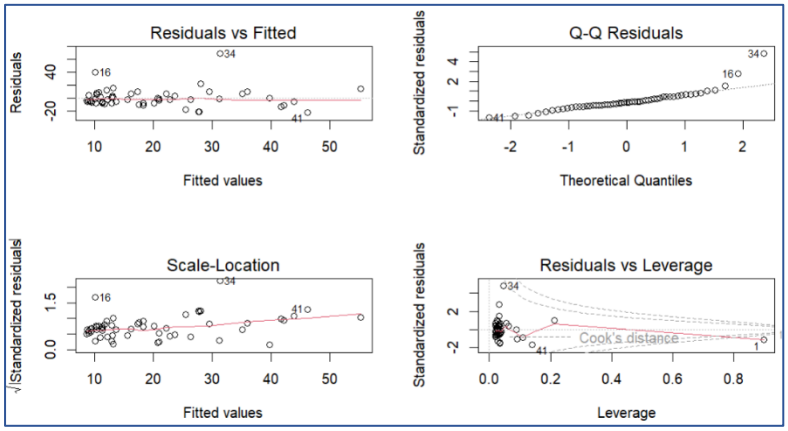
The dependent variable is **life span**, predicted using **body weight** and **gestation period**. Key insights:

- **Intercept (7.58)**: The baseline life span when predictors are zero.
- **Gestation period (p < 0.001)**: Positively and significantly associated with life span, indicating longer gestation predicts higher life span.
- **Body weight (p = 0.245)**: Not statistically significant, suggesting no strong effect on life span.
- **R-squared (0.39)**: Indicates 39% of the variance in life span is explained by the model, a moderate fit.

Overall, gestation period is a significant predictor, while body weight shows minimal influence.

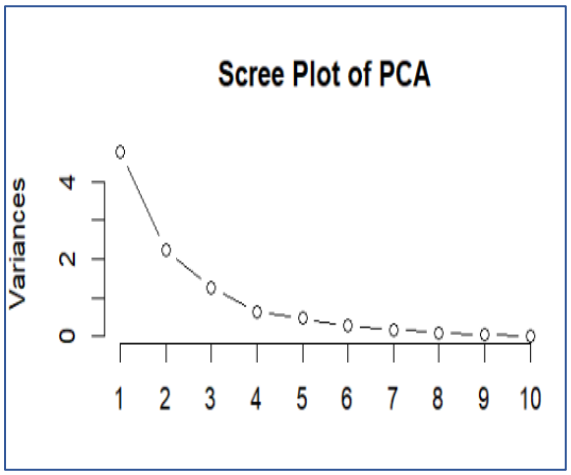
8. Model Diagnostics.

Summary: The residual plots indicate potential violations of assumptions, including heteroscedasticity. Further investigation may be required to ensure model accuracy.



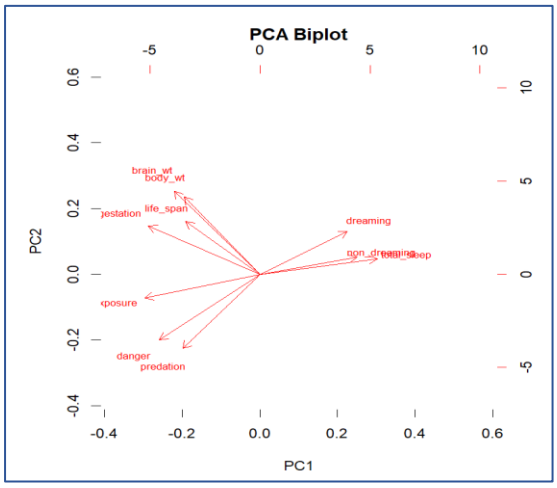
The diagnostic plots indicate potential issues with the model fit. The **Residuals vs. Fitted** and **Scale-Location** plots suggest heteroscedasticity, as residuals show a non-constant variance. The **Q-Q plot** shows significant deviations from normality, particularly in the tails. Influential points (e.g., 34 and 41) are evident in the **Residuals vs. Leverage** plot, which might impact the model's fit. To improve the model, consider transformations (e.g., log), using robust regression, or investigating influential points for their impact. These issues indicate that assumptions of linear regression are not fully met.

9. Principal Component Analysis



The elbow point appears to be at 4. And 90%+ of the variability seems to be explained at around 5 PCs. So 5 number of principal components seems to be an appropriate choice

10. PCA Interpretation



There seems to be a cluster of observations, suggesting moderate similarity in their variable characteristics.Variables like "Body Wt" and "Reproduction" appear to contribute significantly (long arrows). Some variables (e.g., "Dreaming," "Gestation") have similar directions,indicating a positive correlation.Variables pointing in opposing directions (e.g., "Reproduction" vs. "Body Wt") might suggest trade-offs or negative correlations.

Statistical Structures in Data
Prof. Subhajit Dutta

Mohsin Parmar
24BM6JP33

EMPLOYEE ATTRITION

1.Overview of Dataset

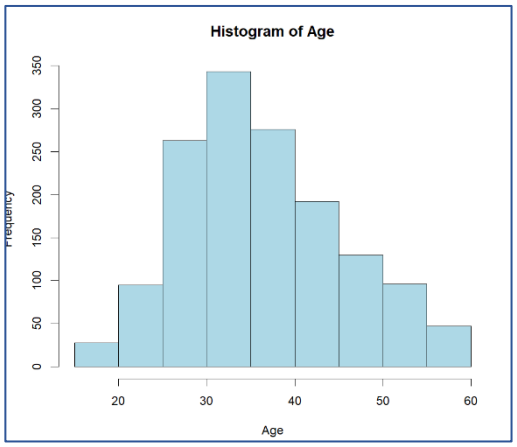
Srno	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education
1	41	Yes	Travel_Rarely	1102	Sales	1	2
2	49	No	Travel_Frequently	279	Research & Development	8	1
3	37	Yes	Travel_Rarely	1373	Research & Development	2	2
4	33	No	Travel_Frequently	1392	Research & Development	3	4
5	27	No	Travel_Rarely	591	Research & Development	2	1

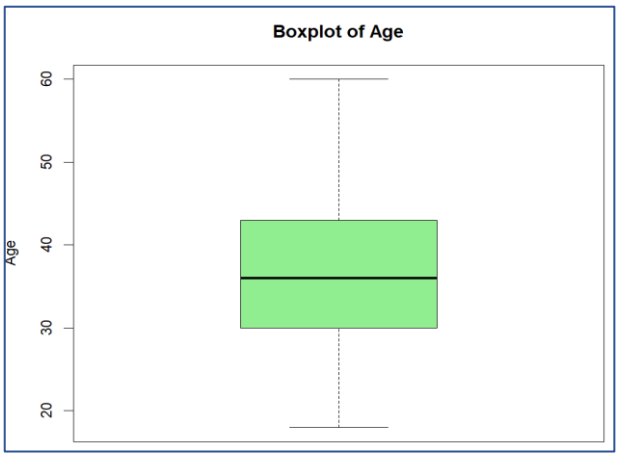
Number of observations: **1470** Number of variables: **35**

2. Summary Statistics of Monthly Income

Mean : **6502.931** Median : **4919** Standard deviation : **4707.957** Minimum : **1009** Maximum : **19999**

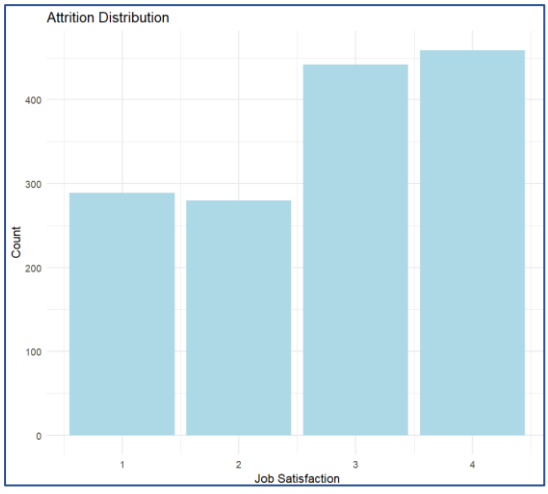
3.Distribution Visualization





The Histogram and the Box plot of Age show that the distribution of Age is right skewed.

4. Categorical Variable Analysis.



The dataset reveals interesting patterns in employee characteristics and job satisfaction. The table shows diverse employee profiles with varying ages (27-49), travel frequencies, and daily rates (279-1392). Research & Development appears to be the dominant department. The attrition distribution graph demonstrates a clear trend in job satisfaction, with higher satisfaction levels (3-4) having notably larger counts, suggesting generally positive employee contentment. Education levels vary from 1-4, and distance from home ranges from 1-8 units. The data indicates that while some employees frequently travel, others rarely do, potentially impacting their work-life balance.

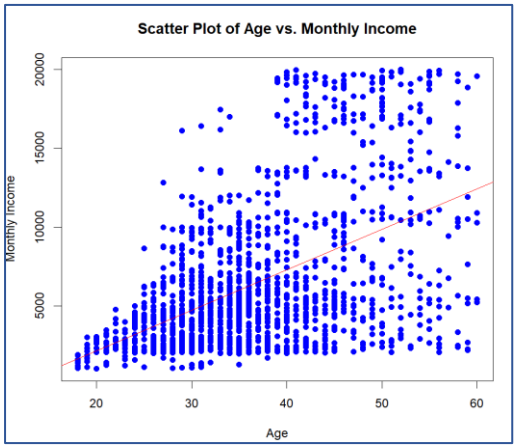
Mohsin Parmar
24BM6JP33

5. Correlation Analysis.

Pearson correlation coefficient between Age and Monthly Income**: 0.4978546**

There seems to be a good correlation between Age and Monthly income. The value is on the lower threshold of **strong positive correlation.**

6. Scatterplot Visualization.



The Scatter plot shows a special group of employees with Age>40 and Monthly Income >15000. They might form a different cluster of employees i.e. " Highly Experienced and Highly Paid Employees".

7.Multiple Linear Regression.

model <- lm(MonthlyIncome ~ YearsSinceLastPromotion + TotalWorkingYears, data = employee_attrition)

	Estimate	Std.	t value	Pr(>|t|)
(Intercept)	1211.32320	137.35967	8.818623	3.217604e-18
YearsSinceLastPromotion	56.03225	26.43013	2.120014	3.417238e-02
TotalWorkingYears	458.26332	10.94611	41.865418	1.104179e-252

	Min	Q1	Median	Mean	Q3	Max	Std_Dev	IQR
25%	-10907.07	-1740.634	-56.47817	1.187261e-14	1402.546	11313.85	2982.803	3143.179

R_Squared	Adjusted_R_Squared
0.5985938	0.5980465

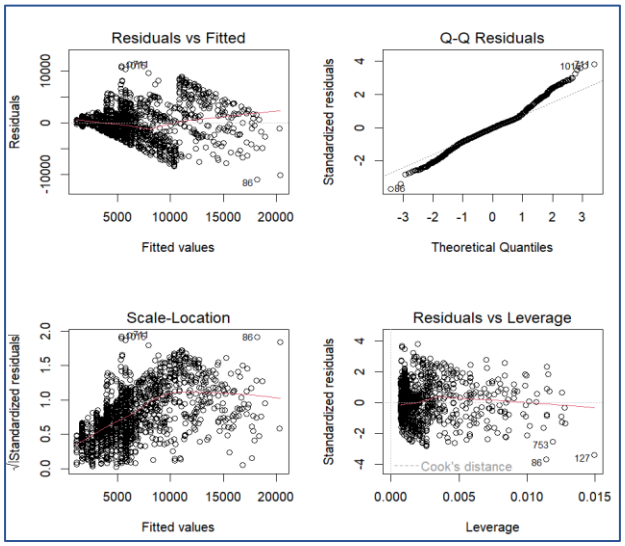
The dependent variable is **monthly income**, predicted using **years since last promotion** and **total working years**. Key insights:

- **Intercept (1211.32)**: The baseline monthly income when predictors are zero.
- **Total working years (p < 0.001)**: Positively and significantly associated with monthly income, indicating that more experience predicts higher income.
- **Years since last promotion (p = 0.034)**: Positively associated with monthly income, though the effect is smaller compared to total working years.

- **R-squared (0.599)**: Indicates 59.9% of the variance in monthly income is explained by the model, a strong fit.

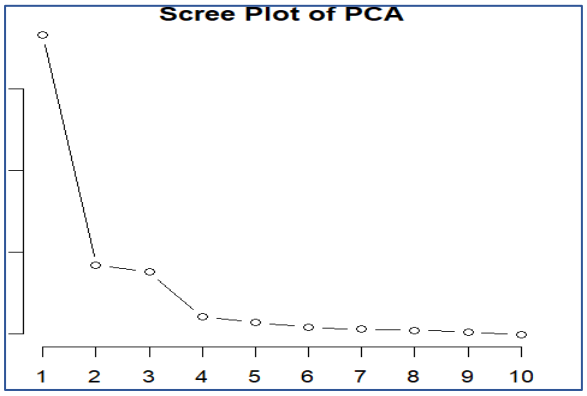
Overall, total working years is a significant predictor, while years since last promotion shows a smaller but still meaningful effect on income.

8. Model Diagnostics.



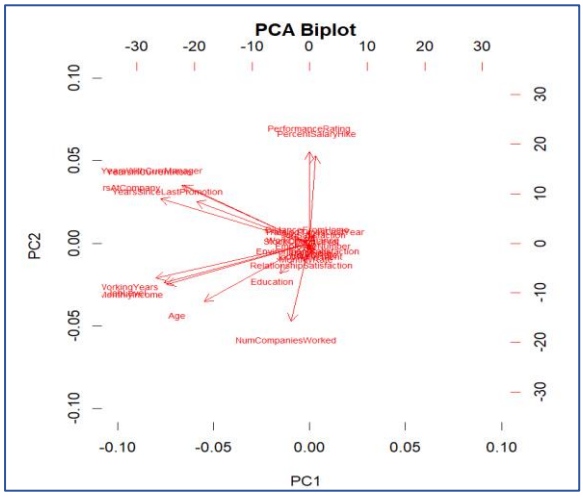
The diagnostic plots reveal important insights about the linear regression model. The Residuals vs Fitted plot shows heteroscedasticity, with increasing spread at higher fitted values, indicating potential violation of constant variance assumption. The Q-Q plot demonstrates relatively normal distribution of residuals, though with some deviation at the tails. The Scale-Location plot confirms non-constant variance, while the Residuals vs Leverage plot identifies potential influential points (Cook's distance). These patterns suggest the model's assumptions aren't fully met, and transformations or robust regression methods might be worth considering for better model fit.

9.Principal Component Analysis.



The scree plot displays the eigenvalues of principal components in descending order. There is a sharp drop from the first to second component, followed by a more gradual decline. The "elbow" appears around component 3, after which the curve flattens significantly. This pattern suggests that the first 3-4 principal components capture most of the data's variance, while subsequent components contribute minimally. For dimensionality reduction purposes, retaining the first three components would likely preserve the most important features of the dataset while effectively reducing its complexity.

10. PCA Interpretation.



The PCA biplot shows relationships between variables, with PC1 and PC2 capturing the most variance. Variables like PerformanceRating, YearsAtCompany, and YearsInCurrentRole are positively correlated, indicating higher performance ratings align with longer tenures. NumCompaniesWorked negatively correlates with Age, suggesting younger employees have worked at fewer companies. JobSatisfaction and EnvironmentSatisfaction are strongly aligned, reflecting a close relationship. The biplot highlights key patterns in employee characteristics and factors influencing attrition or satisfaction.

Statistical Structures in Data
Prof. Subhajit Dutta

Mohsin Parmar
24BM6JP33

WORKERS PRODUCTIVITY

1.Overview of Dataset

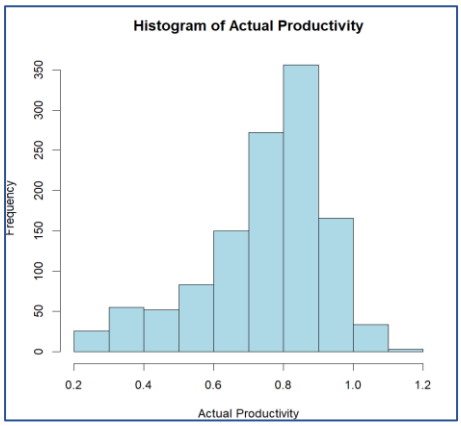
	date	quarter	department	day	team	targeted_	smv	wip	over_time	incentive	idle_men	actual_produ
1	1/1/2015	Quarter1	sweing	Thursday	8	0.80	26.16	1108	7080	98	0	0.9
2	1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94	*NA*	960	0	0	0.8
3	1/1/2015	Quarter1	sweing	Thursday	11	0.80	11.41	968	3660	50	0	0.8
4	1/1/2015	Quarter1	sweing	Thursday	12	0.80	11.41	968	3660	50	0	0.8
5	1/1/2015	Quarter1	sweing	Thursday	6	0.80	25.90	1170	1920	50	0	0.8
6	1/1/2015	Quarter1	sweing	Thursday	7	0.80	25.90	984	6720	38	0	0.8

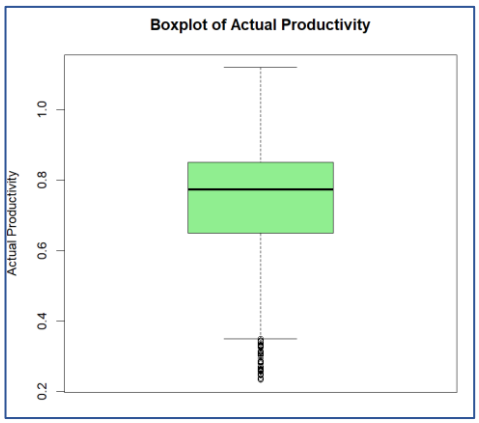
Number of observations: **1197** Number of variables: **15**

2. Summary Statistics of Actual Productivity

Mean: **0.735** Median: **0.773** Standard deviation: **0.174** Minimum: **0.233** Maximum:**1.120**

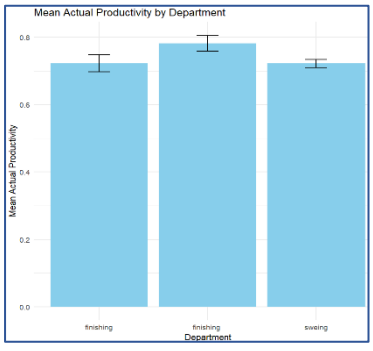
3.Distribution Visualization





The histogram of Actual Productivity seems to be **left skewed**. The box plot of Actual Productivity shows many **outliers on the lower side.**

4.Categorical Variable Analysis



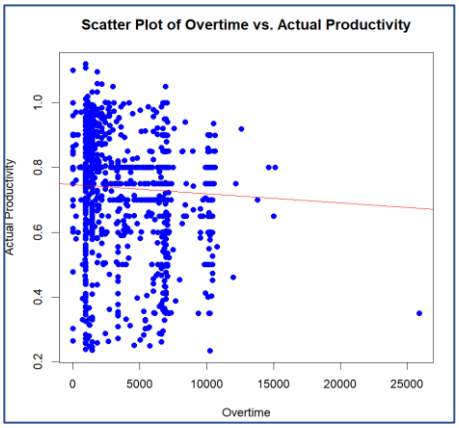
Workers working in the finishing department have a slightly higher mean productivity when compared to workers working in the sewing department.

5. Correlation Analysis.

Pearson correlation coefficient between actual productivity and overtime: **-0.05420584**

The pearson correlation coefficient suggests that actual productivity and overtime are **weakly negatively correlated.**

6. Scatterplot Visualization.



The scatter plot also indicates a weak negative correlation. It also shows a huge outlier for overtime which is sufficiently larger than the other values.

7.Multiple Linear Regression.

model <- lm(actual_productivity ~ smv + over_time + no_of_workers, data = workers_data)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.483649e-01	9.362214e-03	79.9346090	0.000000e+00
smv	-6.559321e-03	1.108483e-03	-5.9173864	4.269068e-09
over_time	-1.072816e-06	2.186249e-06	-0.4907105	6.237214e-01
no_of_workers	2.612662e-03	5.939148e-04	4.3990514	1.184343e-05

	Min	Q1	Median	Mean	Q3	Max	Std_Dev	IQR
25	-	-	0.0444702	-4.66687e-	0.115369	0.378044	0.171667	0.20237

R_Squared	Adjusted_R_Squared
0.03206597	0.02963194

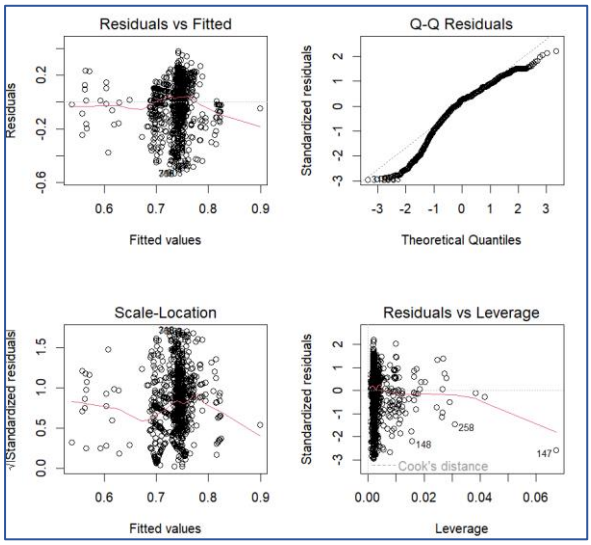
The dependent variable is **actual productivity**, predicted using **SMV**, **overtime**, and **number of workers**. Key insights:

- **Intercept (0.748)**: The baseline productivity when all predictors are zero.
- **SMV (p < 0.001)**: Negatively and significantly associated with productivity, indicating that higher SMV reduces actual productivity.
- **Number of workers (p < 0.001)**: Positively and significantly associated with productivity, suggesting larger teams increase productivity.
- **Overtime (p = 0.623)**: Not statistically significant, implying no strong effect on productivity.
- **R-squared (0.032)**: Only 3.2% of the variance in productivity is explained by the model, indicating a weak fit.

Mohsin Parmar
24BM6JP33

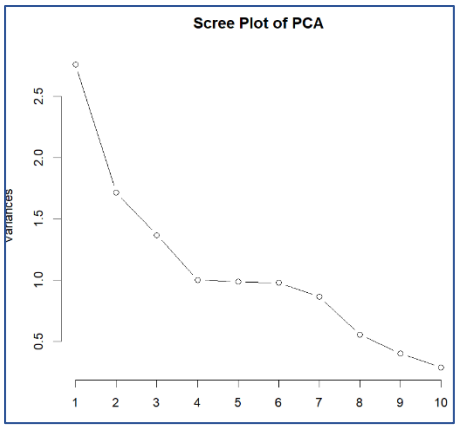
Overall, SMV and the number of workers are significant predictors, with SMV negatively impacting productivity and team size contributing positively. Overtime has minimal influence, and the model explains only a small fraction of productivity variance.

8. Model Diagnostics.



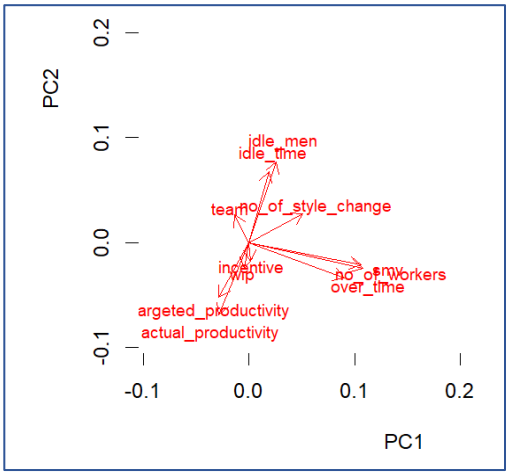
The diagnostic plots indicate issues with the model assumptions. The **Residuals vs Fitted** plot shows non-random patterns, suggesting non-linearity. The **Q-Q Plot** highlights deviations from normality in residuals, particularly at the tails. The **Scale-Location Plot** suggests heteroscedasticity, as variance increases with fitted values. In the **Residuals vs Leverage** plot, points 148, 258, and 147 are influential, as indicated by Cook's distance. These issues imply the model may need refinement, such as transformations or robust regression, to improve fit and meet assumptions.

9. Principal Component Analysis.



The scree plot demonstrates the variance explained by each principal component in descending order. The first component shows the highest eigenvalue at approximately 2.75, followed by a sharp drop to around 1.75 for the second component. A noticeable elbow forms after the third component (around 1.4), where the curve begins to level off significantly. Components 4 through 10 show minimal additional variance contribution, with values below 1.0. This pattern suggests that retaining the first three principal components would capture most of the meaningful variation in the dataset while effectively reducing dimensionality.

10. PCA Interpretation.



The PCA biplot shows relationships between variables, with PC1 and PC2 explaining most of the variance. Variables like **actual_productivity** and **targeted_productivity** are closely aligned, indicating a strong positive correlation. **Idle_time** and **idle_men** cluster together, suggesting these are linked inefficiencies. **No_of_workers**, **overtime**, and **SMV** point in a similar direction, indicating they influence productivity-related variables. **No_of_style_change** is distinct, suggesting minimal correlation with other factors. The biplot helps highlight key drivers of productivity, revealing that certain variables (e.g., idle time and team size) may have opposing effects on operational efficiency.

Statistical Structures in Data
Prof. Subhajit Dutta

Mohsin Parmar
24BM6JP33

GYM DATA TRACKING

1.Overview of Dataset

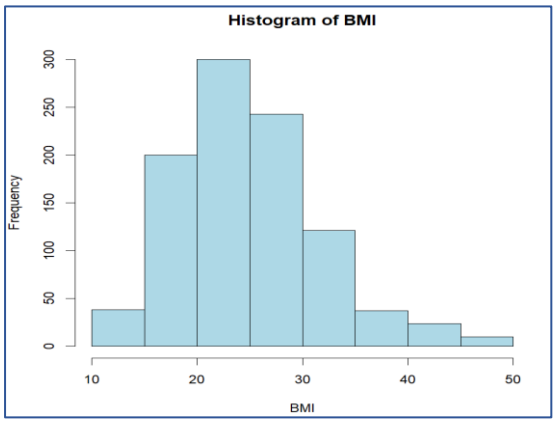
	Age	Gender	Weight	Height	Max	Resting	Session_	Calories	Workout	Fat_	Water_	BMI
1	56	Male	88.3	1.71	180	60	1.69	1313	Yoga	12.6	3.5	30.20
2	46	Female	74.9	1.53	179	66	1.30	883	HIIT	33.9	2.1	32.00
3	32	Female	68.1	1.66	167	54	1.11	677	Cardio	33.4	2.3	24.71
4	25	Male	53.2	1.70	190	56	0.59	532	Strength	28.8	2.1	18.41
5	38	Male	46.1	1.79	188	68	0.64	556	Strength	29.2	2.8	14.39
6	56	Female	58.0	1.68	168	74	1.59	1116	HIIT	15.5	2.7	20.55

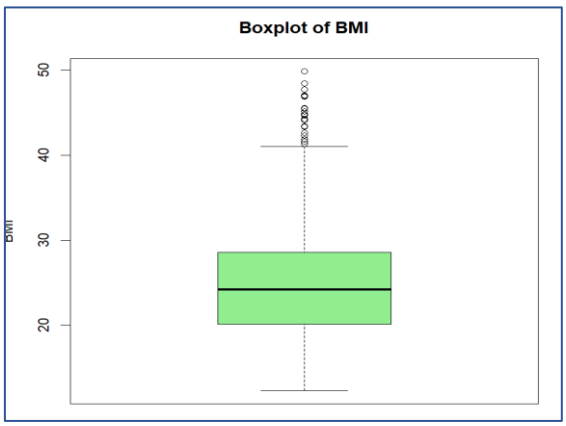
Number of observations: **973** Number of variables: **12**

2. Summary Statistics of BMI

Mean: 24.91 Median: 24.16 Standard deviation: 6.660879 Minimum: 12.32 Maximum: 49.84

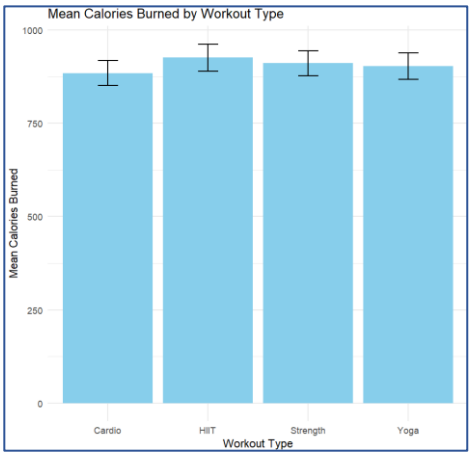
3.Distribution Visualization





Histogram of BMI looks a bit right skewed. The argument is also supported by boxplot which shows outliers on the upper side. The mean of the BMI from the box plot is around 23-24.

4.Categorical Variable Analysis



The mean calories burned is the highest for HIIT(High intensity Interval training) workout type. Followed by Strength, Yoga and Cardio.

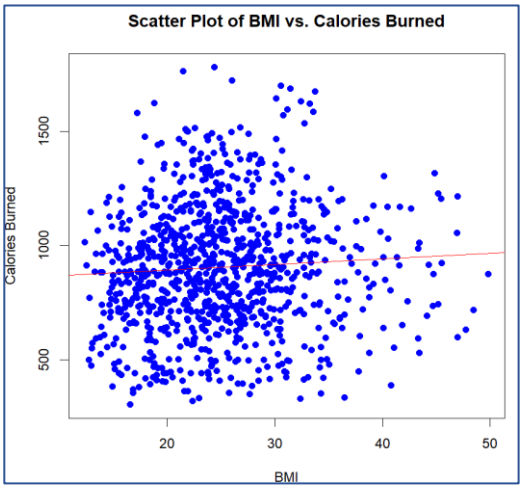
Mohsin Parmar
24BM6JP33

5. Correlation Analysis.

Pearson correlation coefficient between BMI and Calories Burned: **0.05976083**

BMI and Calories Burned are weakly positively correlated.

6. Scatterplot Visualization.



Trend: The scatter plot shows a weak relationship between BMI and calories burned. The red trendline appears nearly flat, suggesting no significant linear correlation between the two variables.

Spread: The data points are widely scattered, indicating high variability in calories burned across different BMI values.

Key Insight: BMI does not seem to strongly influence calories burned based on this data. Other factors might play a more significant role in determining calorie expenditure.

7.Multiple Linear Regression.

model <- lm(Calories_Burned ~ BMI + Age + Weight_in_kg, data = gym_data)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	964.719323	43.8327085	22.009120	2.210210e-87
BMI	-2.868047	2.4796677	-1.156625	2.477105e-01
Age	-3.361778	0.7078047	-4.749585	2.346080e-06
Weight in kg	1.925368	0.7792590	2.470768	1.365347e-02

	Min	Q1	Median	Mean	Q3	Max	Std Dev	IQR
25%	-628.5698	-174.5382	-5.674461	4.202188e-15	168.8244	822.8252	268.0582	343.3625

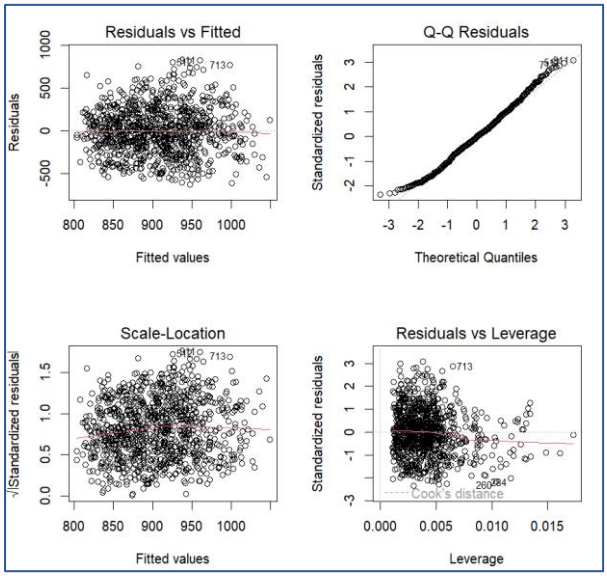
	R_Squared	Adjusted_R_Squared
1	0.03333882	0.03034606

- **Intercept (964.72)**: The baseline value of Calories_Burned when all predictors (BMI, Age, Weight_in_kg) are zero.
- **BMI (-2.87)**: Not statistically significant ($p = 0.247$), suggesting BMI does not strongly influence the number of calories burned in this model.
- **Age (-3.36)**: Statistically significant ($p < 0.001$), indicating that as age increases, the number of calories burned decreases.

- **Weight_in_kg (1.93)**: Statistically significant (p = 0.014), suggesting that as weight increases, calories burned also increase.
- **R-Squared (0.033)**: The model explains only about 3.3% of the variance in calories burned, indicating a weak fit.

Conclusion: While Age and Weight_in_kg show significant relationships with calories burned, the model as a whole does not explain much of the variance in the data.

8. Model Diagnostics.



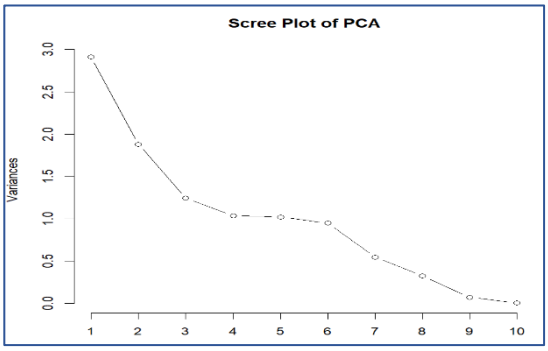
Linearity & Homoscedasticity (Residuals vs. Fitted): Residuals are randomly scattered, supporting linearity. Slight variance inconsistency hints at mild heteroscedasticity.

Normality (Q-Q Plot): Residuals mostly align with the diagonal line, indicating approximate normality, though minor outliers are present.

Variance Consistency (Scale-Location): Residuals show consistent spread, with slight deviations suggesting mild heteroscedasticity.

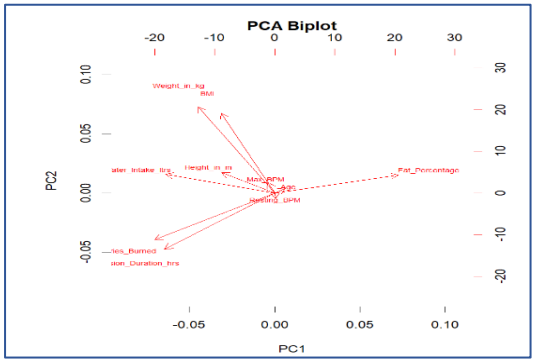
Influential Points (Residuals vs. Leverage): A few data points (e.g., 713, 268) near Cook's distance lines could significantly affect the model and warrant further investigation.

9. Principal Component Analysis.



The first 8 principal components explain approximately 90% of the variability.This suggests that using the first 8 principal components would be sufficient to represent the data while maintaining most of its important features, as subsequent components contribute minimal additional variance explanation

10. PCA Interpretation.



Weight, BMI, and height cluster together, indicating a strong positive correlation. Fat percentage aligns closely with this group. Sleep-related variables (sleep_duration, sleep_efficiency) form another cluster, suggesting their interrelation. Sedentary minutes and steps show an inverse relationship, as expected.