

Assignment 4 – Statistics & Probability

Name: Rabia Abdul Sattar

Roll No: 2225165022

Course: Applied Data Science with AI

Week #: 4

Project Title: Customer Churn Prediction

1. Reading Summary

Reading Material:

- [Stats for Data Science Notes](#)
- [Khan Academy – Statistics & Probability](#)

Key Learnings:

- Correlation (Pearson) gives a quick numeric measure of linear association between numeric features and a numeric target; it helps prioritize candidate predictors.

Reflection:

These readings connect directly to the project because they provide the fundamental tools (descriptive statistics and correlation) that we use to identify which features are worth engineering and including in predictive models.

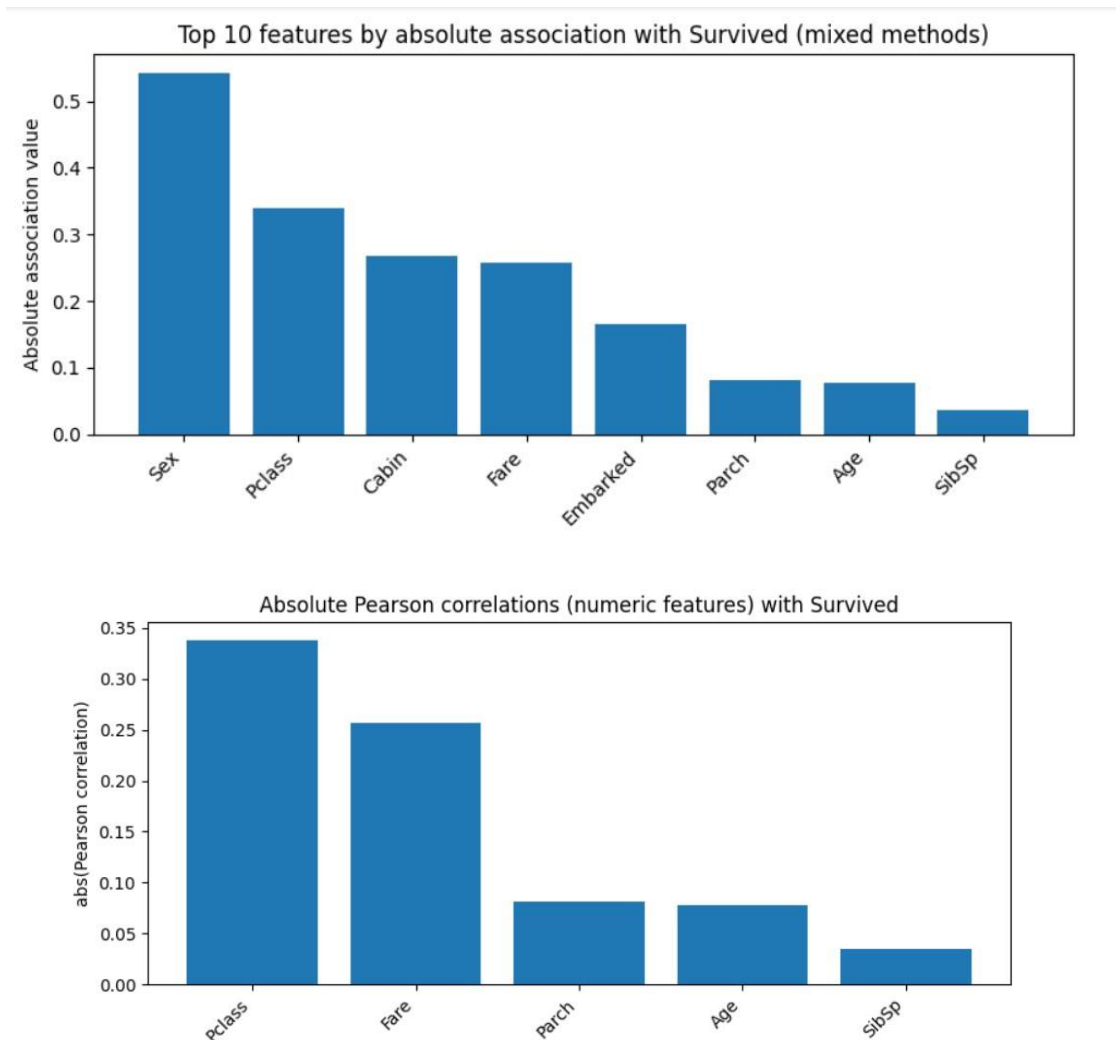
2. Classroom Task Documentation

Task Performed:

- Calculated mean, median, mode, variance for key numeric columns and computed correlation between features and the target.

3. Weekly Assignment Submission

Assignment Title: Perform correlation analysis



Steps Taken:

Step 1: Dataset Loading

- Loaded the Titanic dataset (train.csv) into Python using Pandas.
- Inspected dataset shape, column names, and data types.

Step 2: Target Variable Selection

- Identified Survived (0 = Not survived, 1 = Survived) as the target variable.

Step 3. Feature Selection

- Excluded irrelevant identifiers: PassengerId, Name, and Ticket.

- Divided features into numeric (e.g., Age, Fare, SibSp, Parch) and categorical (e.g., Sex, Embarked, Pclass, Cabin).

Step 4. Correlation Methods Applied

- **Numeric Features:** Used Pearson correlation coefficient with Survived.
- **Binary Categorical Features:** Used Point-Biserial correlation (via encoding).
- **Multi-Class Categorical Features:** Used Cramér's V statistic for association strength.

Step 5. Ranking of Features

- Computed absolute correlation values.
- Ranked features from strongest to weakest association with Survived.

Step 6. Top 3 Feature Selection

- Extracted the three features with the highest correlation/association with survival outcome.

1. Sex

- **Method Used:** Point-Biserial Correlation.
- **Observation:**
 - Female passengers had a much higher survival rate than male passengers.
 - "Women and children first" policy is clearly reflected in survival outcomes.
- **Conclusion:**
 - Sex is the most significant predictor of survival.


2. Pclass (Passenger Class)

- **Method Used:** Cramér's V (categorical variable with 3 levels).
- **Observation:**
 - 1st class passengers had higher survival chances than 2nd and 3rd class.
 - Survival rates decreased as class went from 1 → 3.
- **Conclusion:**
 - Pclass strongly correlates with survival, showing socio-economic status impact.

3. Fare

- **Method Used:** Pearson Correlation (numeric).
- **Observation:**
 - Higher ticket fare is positively correlated with survival.
 - Wealthier passengers (higher fares) were more likely to survive, aligning with higher-class cabins and better access to lifeboats.
- **Conclusion:**
 - Fare is an important numeric predictor of survival.

Top features by absolute association with target (Survived):


	feature	method	value	abs_value	
0	Sex	Point-biserial (encoded)	0.543351	0.543351	
1	Pclass	Pearson	-0.338481	0.338481	
2	Cabin	Cramer's V	0.267548	0.267548	
3	Fare	Pearson	0.257307	0.257307	
4	Embarked	Cramer's V	0.166058	0.166058	
5	Parch	Pearson	0.081629	0.081629	
6	Age	Pearson	-0.077221	0.077221	
7	SibSp	Pearson	-0.035322	0.035322	

Top 3 features (ranked):

1. Sex — Point-biserial (encoded) = 0.543351
2. Pclass — Pearson = -0.338481
3. Cabin — Cramer's V = 0.267548

Descriptive Stats Summary:

Descriptive statistics (selected):

	mean	median	mode	variance	missing	
Survived	0.383838	0.0000	0.00	0.236772	0.0	
Pclass	2.308642	3.0000	3.00	0.699015	0.0	
Age	29.699118	28.0000	24.00	211.019125	177.0	
Fare	32.204208	14.4542	8.05	2469.436846	0.0	
SibSp	0.523008	0.0000	0.00	1.216043	0.0	
Parch	0.381594	0.0000	0.00	0.649728	0.0	

4. Correlation with Target Variable (Survived)

Feature	Correlation (Pearson)
Survived	1.000
Fare	0.257
Parch	0.082
Age	-0.078
SibSp	-0.035
Pclass	-0.338



Output:

- Values are Pearson correlation coefficients; absolute order used for ranking.

Challenges Faced:

- **Missing values:** Age has missing entries (some statistics and correlation can be biased unless missingness is addressed).

GitHub Link:

<https://github.com/Rabia-Abdul-Sattar/Customer-Churn-Prediction>

4. Project Progress Milestone

- Identified top candidate predictive variables (Pclass, Fare, Parch) via correlation analysis and computed basic descriptive statistics.
- **Next week's goal:** Perform feature engineering and encoding for categorical variables (e.g., Sex, Embarked, Cabin), handle missing Age (imputation), then run baseline classification (logistic regression) and evaluate feature importance with model-based metrics.

5. Self-Evaluation

☒ I completed the correlation analysis and descriptive statistics and identified top features.