**A report covering my approach on each Dataset**

In the Data Science internship carried out, I worked with given be LogicLeap. I worked with the Titanic Dataset, Airline passenger dataset and A/B testing dataset. The goal was to apply analytics and statistical method using python to derive insights into the dataset, interpret and also predict where necessary.

1. **Titanic Dataset Analysis**

In the titanic dataset, the columns consist of sex, age, class, embarked, survived etc. The target variable is the survived. It has the value of “0” and “1”. In order to carry out analysis in the dataset, it is necessary to clean it up and handle the missing data, either by filling with median or mode for numerical columns and forward or backward filling, or replacing with an “Unknown” value for categorical column. Also, categorical encoding was carried out on the columns. By using Label encoding, dummies or binary encoding. We were able to successfully connect the values to numerical values.

Analysis was carried out by plotting charts and graphs across different variables to depict the relationship between each variable. Since survival is the target variable, the visualizations are all in relation to the “Survived” variable.

**Challenges:** The dataset contained some degree of missing data. But since the data was cleaned and we didn’t model the dataset, there was not a lot of challenge in that area.

The number of non-survivors is more than the survivors, which made the data imbalanced.

**Outcomes:** It is seen that in the dataset, according to visualizations, females and people in higher class of the titanic were seen to survive more. The lower class and young adults in their late 20’s had the most mortality rate. Demographic and class segmentation can influence survival in emergencies, guiding safety planning and targeted resource allocation.

1. **Airline Passengers (SARIMA Forecasting)**

The airline passengers dataset contains the date and number of passengers for each month. In this project, there was need to carry out time series decomposition, to understand the trend, understand how the data reacts in each season. The decomposition showed increase in trend with time and also the data was non-stationary using, until seasonal differencing was used and it is seen that the data is stationary.

The SARIMA model was seen to give better accuracy of the prediction of the dataset. With an R2-score of 91%, we were able to see how well the model performed. The comparism was carried out using a visualization tool.

**Challenges:** One of the challenges faced was that the data needed differencing before we could achieve stationarity. Different approaches were carried out until the seasonal differencing looked to be the best.

Model selection was also a bit of a challenge. Only the SARIMA was used in the jupyter file, but ARIMA was tested also and seen to be of lower accuracy.

**Outcomes:** Airlines can use SARIMA forecasts to plan resource allocation.

1. **A/B Testing Dataset Statistical Analysis**

The A/B testing dataset contains the user id, time stamp, landing page, group and converted. This is where the null and alternative hypothesis, my null hypothesis is that the landing pages have no effect on the converted. Meaning new or old page changes, does it mean more people convert in the landing page by performing any actions?

The Z-test was carried out here for the statistical analysis because it is necessary when we have large dataset and also because it is between two category values to check for significant difference. The test is also a two-tailed test.

**Challenges:** Interpreting p-values and ensuring business context alignment with statistical significance and also ensuring proper random assignment and ruling out confounding factors.

**Outcomes:** The p-value from hypothesis testing was greater than 0.05, meaning no statistically significant difference between the old and new landing pages. Business recommendation maintain the current version until stronger evidence supports switching.