Capstone Project 1: Statistical Data Analysis

Think of the following questions and apply them to your dataset:

- Are there variables that are particularly significant in terms of explaining the answer to your project question?

- Are there significant differences between subgroups in your data that may be relevant to your project aim?

- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

- What are the most appropriate tests to use to analyze these relationships?

Submission: Write a 1-2 page report on the steps and findings of your inferential statistical analysis. Upload this report to your GitHub and submit a link. Eventually, this report will get incorporated into your milestone report.

Project goal:

The ecommerce company would like to predict lifetime value or revenue for customers acquired through the marketing organization. In the project, we are asked to predict 2-year customer lifetime value (24m LTV) based on the first purchase information. The prediction will help marketing organization to allocate marketing spend and create proper marketing strategies.

From the last exercise (data story-telling), we knew that Variable revenue\_24 is highly-skewed and follows long-tailed distribution. So let’s do a log-transformation and check the distribution again.

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

By comparing two plots, log-transformation is helpful and transforms the skewed data closer to bell-curve.

I applied same log-transformation on first\_purchase\_revenue and created a variable log10\_first\_value.

Let’s check the relationship between two numeric variables

A close up of a map

Description automatically generated

Looking at this plot, we can easily say: As first purchase revenue increases, corresponding 24month lifetime value automatically increasing in general. Therefore, they are positively correlated with each other. Certainly there are many outliers as well in our dataset that we can see at top left, where purchase revenue is low but corresponding 24month lifetime value is extremely high. Also from the density plots on top and right, the distribution of two variables are very similar.

Another interesting finding here, line is almost diagonal which means the coefficient is very close to 1.

Let’s run a linear regression to check coefficient.

A screenshot of a social media post

Description automatically generated

The Coef = 1.0427! which is very close to 1. And the R-squared is very large as 0.988, meaning this variable is able to explain 98.8% of the data. From the business side, we know about 70% of the customers only purchase once during the 24month lifetime period. However, the log-transformation reduced the value range of the variables which increases bias. Let’s do another linear analysis without log transformation.

A screenshot of a cell phone

Description automatically generated

From the results above, the coef= 1.37 which is higher than log-transformed one. R-squared =0.425 which first purchase variable is able to explain 43.5% of the data set. This is telling us that the first purchase variable is very important and we’ll definitely include it when doing modeling.

Now, let’s look at categorical variables.

**Marketing channels:**

A screenshot of a cell phone

Description automatically generated

From the summary above, we noticed that about 14.5% of the customers completed their first purchase via Direct channel while 8.2% of customers via Brand SEM channel. This is very interesting because customers know your brand name and they are able to find you via either direct channel or brand search (brand SEM). So we are wondering is that a real difference or just data collection/sample bias.

Let’s conduct a t-test.

1. Frame hypothesis: H0: mean of ltv from direct channel = mean of ltv from brand SEM; H1: mean of ltv from direct channel =! mean of ltv from brand SEM; statistical significance level = 95%
2. Doing a t-test:

A screenshot of a cell phone

Description automatically generated

p-value =0.01 < 0.05, we reject H0. And we could let marketing team know that they should treat customers differently from these two channels.

Let’s look at all paid marketing channels where the dollars spent.

We conduct ANOVA test for three paid marketing channels:

A screenshot of a cell phone

Description automatically generated

p-value is very small, telling us these channels are different. But we don’t know which one is different from another or they are all different. So let’s do t-test.

A screenshot of a social media post

Description automatically generated

p-value are very very small, telling us all three channels are significantly different from each other and marketing team should treat those customers separately.

**Shipping address type:**

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

We have 60% of Residential type customers VS 30% of Business type customers. However, Business type customers generate 46% of the total revenue while Residential type customers generate 42% of the total revenue. This is telling us that business type customers submitting large orders which is very good to the business. Let’s plot violin plot to check our guessing.

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Clearly, business type customers has higher mean and less variance which is more robust compared to residential type customers. Also, let’s run t-test:

A screenshot of a cell phone

Description automatically generated

p-value = 0! They are significantly different.