Data Intake Report

Name: <Insight for Cab Investment firm>

Report date: <01.03.2021>

Internship Batch:< LISP01>

Version:<1.0>

Data intake by:<Irem Tanrıverdi>

Data intake reviewer:<intern who reviewed the report>

Data storage location: < https://github.com/DataGlacier/DataSets.git>

**Tabular data details:**

**Name :Cab\_Data.csv**

|  |  |
| --- | --- |
| **Total number of observations** | <359392> |
| **Total number of files** | <1> |
| **Total number of features** | <8> |
| **Base format of the file** | <.csv> |
| **Size of the data** | <21,2 MB> |

**Name :City.csv**

|  |  |
| --- | --- |
| **Total number of observations** | <20> |
| **Total number of files** | <1> |
| **Total number of features** | <3> |
| **Base format of the file** | <.csv> |
| **Size of the data** | < 4 KB> |

**Name : Customer\_ID.csv**

|  |  |
| --- | --- |
| **Total number of observations** | <49171> |
| **Total number of files** | <1> |
| **Total number of features** | <4> |
| **Base format of the file** | <.csv> |
| **Size of the data** | < 1,1 MB> |

**Name :Transaction\_ID.csv**

|  |  |
| --- | --- |
| **Total number of observations** | <440098> |
| **Total number of files** | <1> |
| **Total number of features** | <3> |
| **Base format of the file** | <.csv> |
| **Size of the data** | <9 MB> |

**Name :US\_bank\_holidays.csv**

|  |  |
| --- | --- |
| **Total number of observations** | <89> |
| **Total number of files** | <1> |
| **Total number of features** | <3> |
| **Base format of the file** | <.csv> |
| **Size of the data** | <4 KB> |

**Proposed Approach:**

* Mention approach of dedup validation (identification)

"Duplication" means that if we have repeated data in our dataset. This could be due to things like data entry errors or data collection methods. To take a look at how many rows are duplicated in our dataset and where they are, I visualized the duplication. To plot duplicates, first I created a data frame with a logical vector indicating whether or not a specific row is duplicated elsewhere in the dataset and a numeric vector of the index of each row using “duplicated” function in R. Then, I plotted that information so that each duplicated row shows up as a black line. Then, using this data frame, I looked at which rows are duplicated and number of duplicated rows using “count” commend in R.

* Assumptions :

**Tests used:**

1. **T-test**

It is a test for comparing means of two sample. T-test has three assumptions. First one is samples must normally distributed. Second one is variance of two sample must be homogeneous. Third one is samples must be independent. In this project normality of sample which were applied the t-test did not met. To satisfy normality, some transformation like Box-cox transformation were applied, but normality still not met. Thus, non-parametric version of the t-test which is called Wilcoxon Rank Sum test was applied.

1. **Two-way ANOVA**

A two-way ANOVA is used to estimate how the[mean](https://www.scribbr.com/statistics/mean/) of a [quantitative variable](https://www.scribbr.com/methodology/types-of-variables/) changes according to the levels of two categorical variables. In this project, it was used to examine how two independent variables, in combination, affect a dependent variable. Its assumptions are :

* Dependent variable should be continuous.
* Two independent variables should be in categorical.
* Sample independence – that each sample has been drawn independently of the other samples.
* Variance Equality – That the variance of data in the different groups should be the same
* Normality – That each sample is taken from a normally distributed population

Variable used in this analysis again did not satisfied the normality, so for this analysis normality was assumed and continued analysis with this assumption.

1. **Chi-square independence test**

* The data in the cells should be frequencies or counts of cases rather than percentages or some other transformation of the data.
* The levels (or categories) of the variables are mutually exclusive. That is, a particular subject fit into one and only one level of each of the variables.
* Each subject may contribute data to one and only one cell in the χ2.
* The study groups must be independent.
* There are 2 variables, and both are measured as categories, usually at the nominal level.

For this test all assumptions are satisfied.

1. **Logistic regression**

* The dependent variable should be binary.
* Observations must be independent of each other.  In other words, the observations should not come from repeated measurements or matched data.
* Independent variables should not be too highly correlated with each other.
* Logistic regression assumes linearity of independent variables and log odds.
* Logistic regression typically requires a large sample size.

Linearity of independent variables and log odds was assumed. Moreover, for this model some variables are highly correlated. Thus, another model was tried.