A · P · U

ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Module Code	:	CT075-3-2-DTM – DATA MANAGEMENT
Intake Code	:	APU2F2402CS(DA)/APD2F2402CS(DA)
Lecturer Name	:	DR. MURUGANANTHAN VELAYUTHAM
Hand in Date	:	29TH MAY 2024, WEDNESDAY, 11:59PM
Lab No.	:	LAB 1
Group No.	:	GROUP T
Group Leader	:	CHURILOV MIKHAIL

Student ID	**Student Name**
TP067853	AZELEA GLORY NG ZI-LIN
TP072847	CHURILOV MIKHAIL
TP074470	JANE LEE
TP068019	RAVIN A/L KANAGARAJAN

Table of Contents

1.0 Introduction

Effective data management has become critical in today's environment of data-driven decision-making. Data management guarantees compliance to organizational norms, regulations, and policies while greatly improving the quality of business-related decisions and assisting in data optimization. This assignment explores the field of transportation, emphasizing the complex procedures associated with data pre-processing and the use of data mining methods for obtaining valuable information.

On a daily basis, the transportation industry produces huge amounts of data due to its complexity and dynamic nature. This data, which includes passenger information, logistical details, and vehicle telemetry, has the power to transform operations, boost productivity, and raise satisfaction among consumers. To convert the raw data into a format that can be analyzed and relied upon, however, strong pre-processing methods are required because it is frequently noisy, inconsistent, and incomplete.

In order to prepare the data for analysis, we will examine a transportation dataset in this assignment and apply a variety of data pre-processing techniques. After that, we'll use SAS Studio to apply an appropriate data mining technique to find trends and insights that help guide transportation-related decision-making. Through this process, we aim to highlight the critical role of data pre-processing in improving model performance and the overall analytical outcomes.

2.0 Data Dictionary

Attribute	Description	Type
MEMBER_NO	Membership Card Number	Nominal
FFP_DATE	Enrolment Date	Interval
FIRST_FLIGHT_DATE	First Flight Date	Interval
GENDER	Gender	Nominal
FFP_TIER	Membership Tier	Ordinal
WORK_CITY	Work City	Nominal
WORK_PROVINCE	Work Province	Nominal
WORK_COUNTRY	Work Country	Nominal
AGE	Age	Ratio
LOAD_TIME	Observation window end time	Interval
FLIGHT_COUNT	Number of flights	Ratio
BP_SUM	Total basic points in observation window	Ratio
SUM_YR_1	Total fare for the first year	Ratio
SUM_YR_2	Total fare for the second year	Ratio
SEG_KM_SUM	Total flight distance in observation window	Ratio
LAST_FLIGHT_DATE	Last flight date	Interval
LAST_TO_END	Time from last flight to observation window end	Ratio
AVG_INTERVAL	Average time interval between flights	Ratio
MAX_INTERVAL	Maximum time interval between flights in observation window	Ratio
EXCHANGE_COUNT	Number of points exchanges	Ratio
avg_discount	Average discount rate	Ratio
Points_Sum	Total cumulative points	Ratio
Point_NotFlight	Number of non-flight point changes	Ratio

3.0 Exploratory Data Analysis (EDA) and Data Pre-Processing

3.1 MEMBER_NO

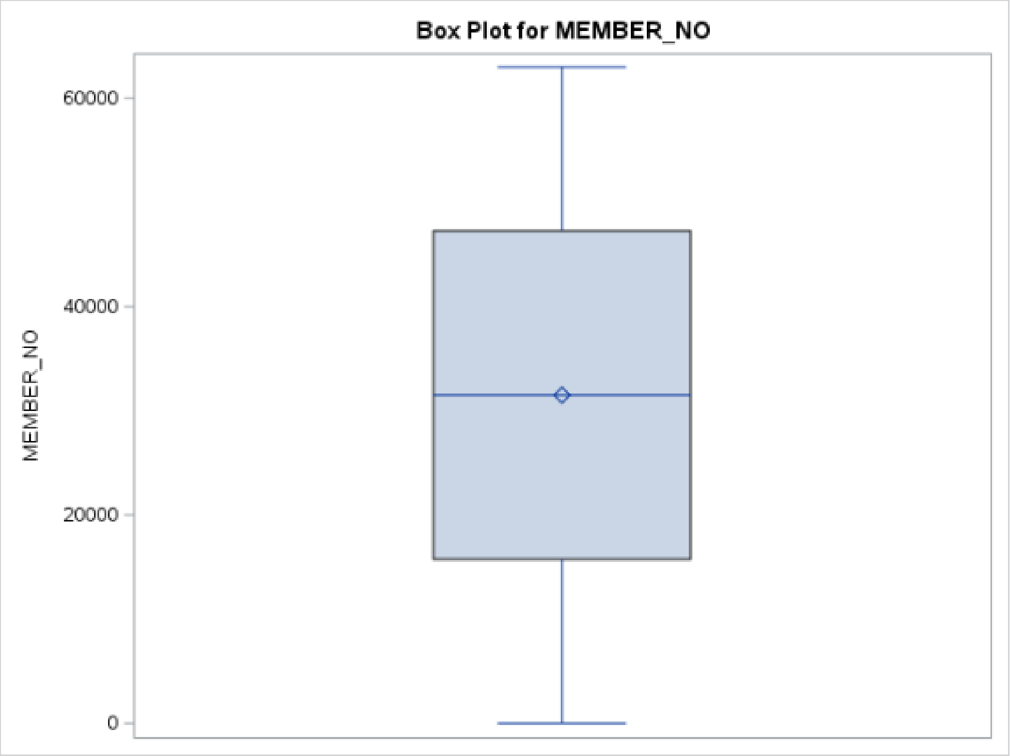


Figure 3.1.1: Box Plot of MEMBER_NO

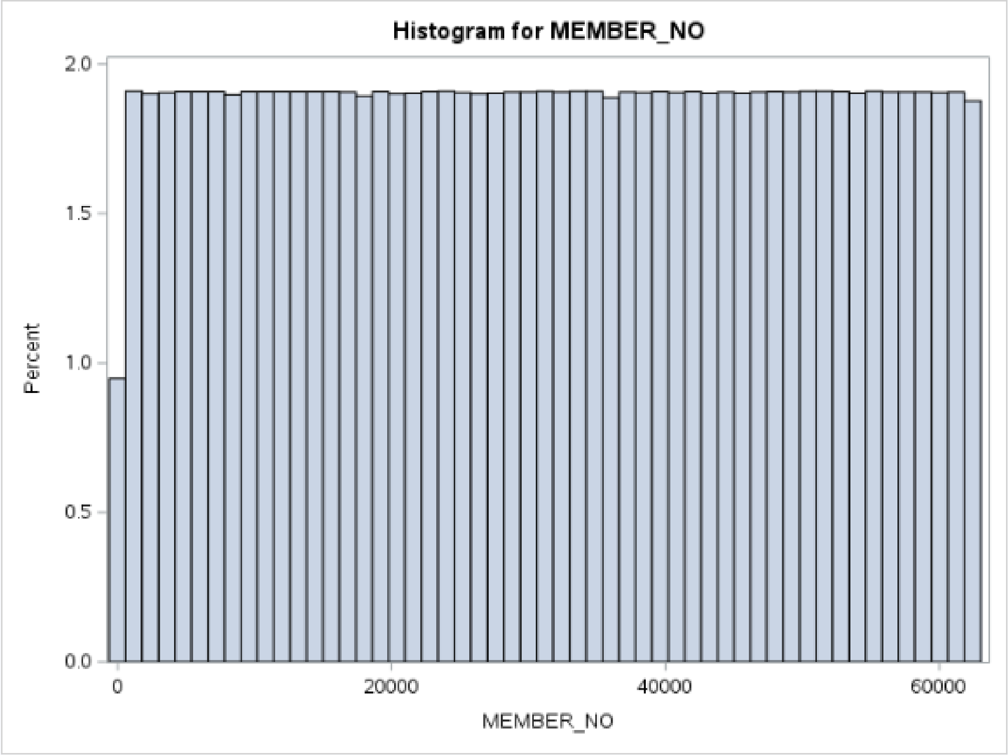
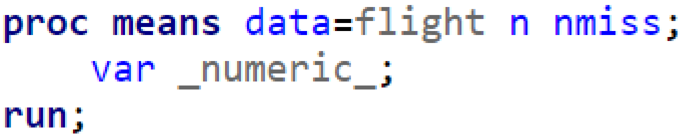


Figure 3.1.2: Histogram of MEMBER_NO

Pre-Processing Methods:

Cleaning Method: Removing Duplicates



```
proc means data=flight n nmiss;
    var _numeric_;
run;
```

Figure 3.1.3

The first step in pre-processing MEMBER_NO was to identify and remove any duplicate entries. Duplicates can distort analysis and lead to inaccurate conclusions. Using SAS, we applied the following code to eliminate duplicates. This ensured that each MEMBER_NO is unique, preserving the integrity of the membership data.

3.2 FFP_DATE

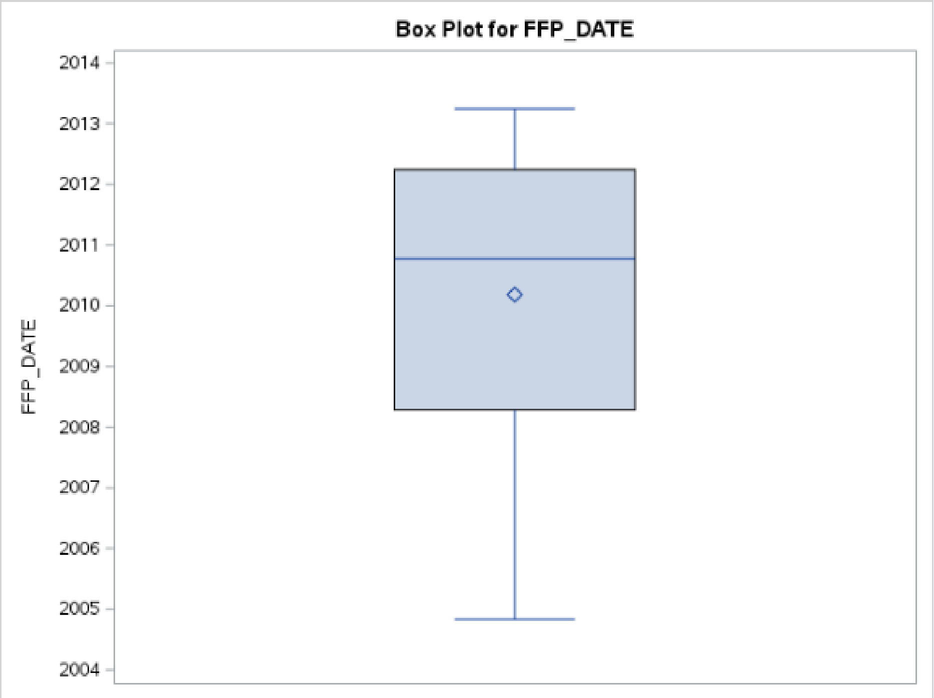


Figure 3.2.1: Box Plot of FFP_DATE

The box plot of FFP_DATE shows that there are no significant outliers exist in this data. The data distribution is relatively symmetrical, with the interquartile range covering the years from 2007 to 2012. The whiskers have extended has cover the full range of the data points from 2004 to 2013.

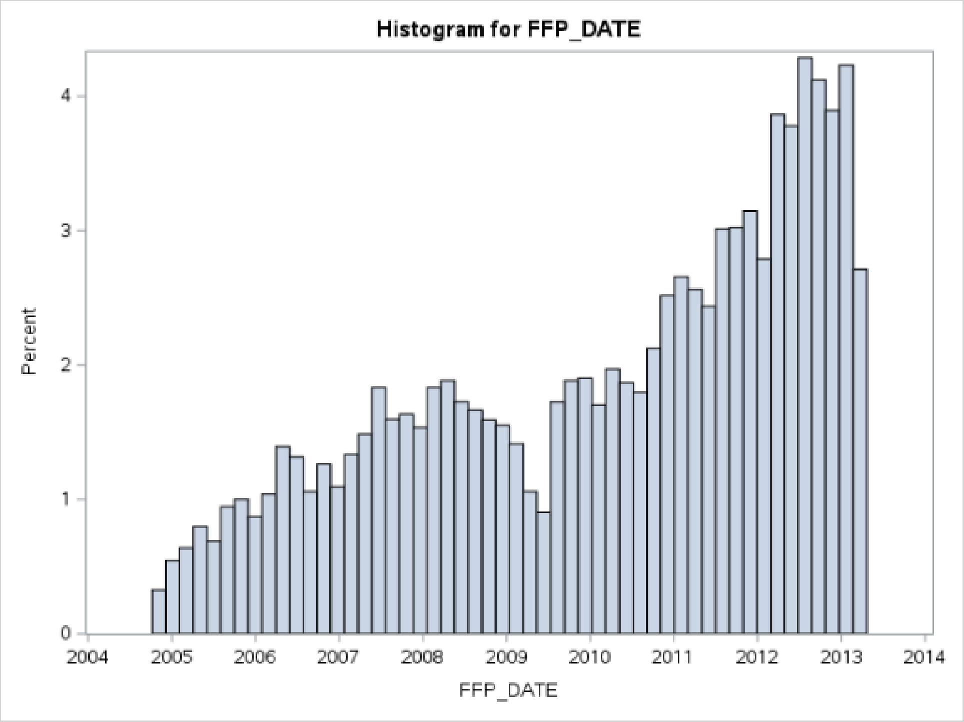


Figure 3.2.2: Histogram of FFP_DATE

In this histogram of FFP_DATE, it shows a fairly even distribution of values with a slightly increase in frequency years by years. This can determine that there are more customers or passengers have joined the fly program in recent year, especially around 2013. The histogram confirms that there are no extreme outliers existed.

Pre-Processing Method:
Cleaning Method: Checked for missing values.
In FFP_DATE, the data doesn’t exist any missing values and the data distribution is clean and consistent, there are no pre-processing was required for FFP_DATE.

3.3 FIRST_FLIGHT_DATE

Before Pre-Processing:

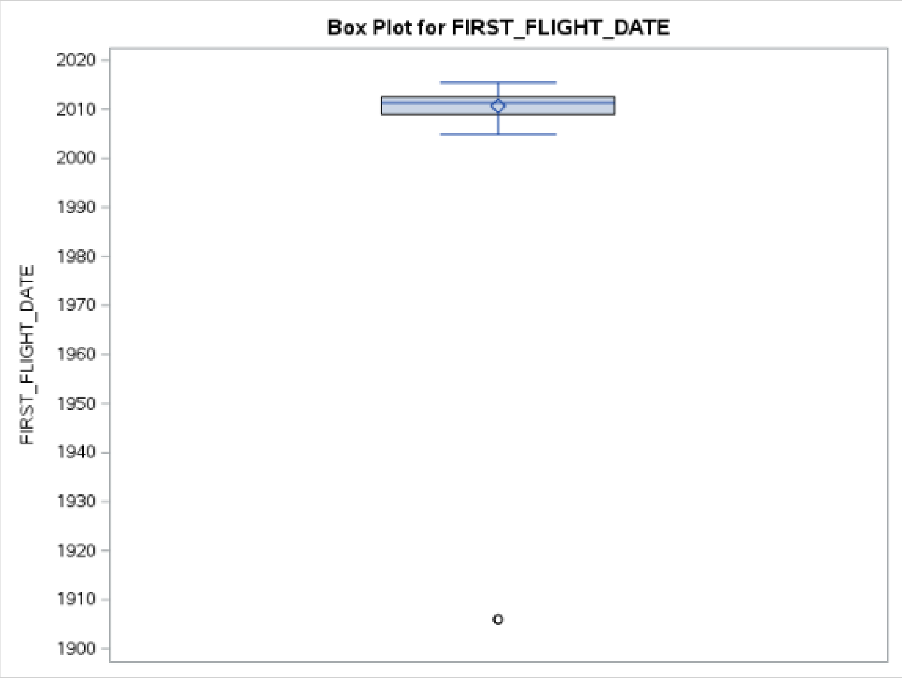


Figure 3.3.1: Box Plot of FIRST_FLIGHT_DATE

The box plot reveals an outlier significantly earlier than the bulk of the data points, which the box plot concentrates between the years of 2009 and 2012. There are a single outliers existed in around 1909, which indicates a data entry anomaly.

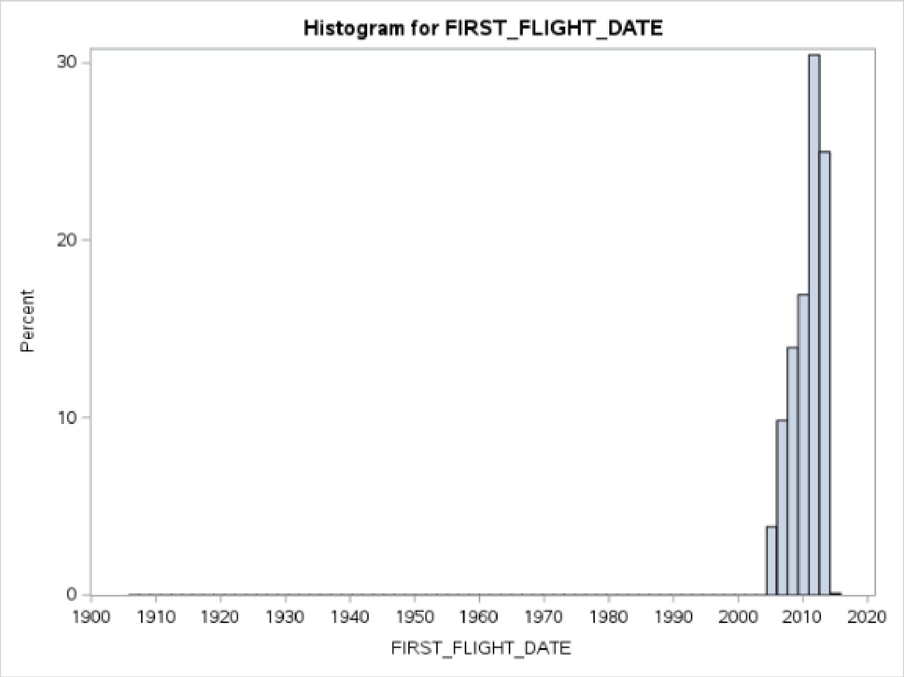


Figure 3.3.2: Histogram for FIRST_FLIGHT_DATE

The histogram shows a normal distribution of FIRST_FLIGHT_DATE with a peak around 2012 to 2013. This can confirm the central tendency of the data is around this period.

Pre-Processing Method:
Since there I only one anomaly, it can be removed manually to maintain the integrity of the dataset.

After Pre-Processing:

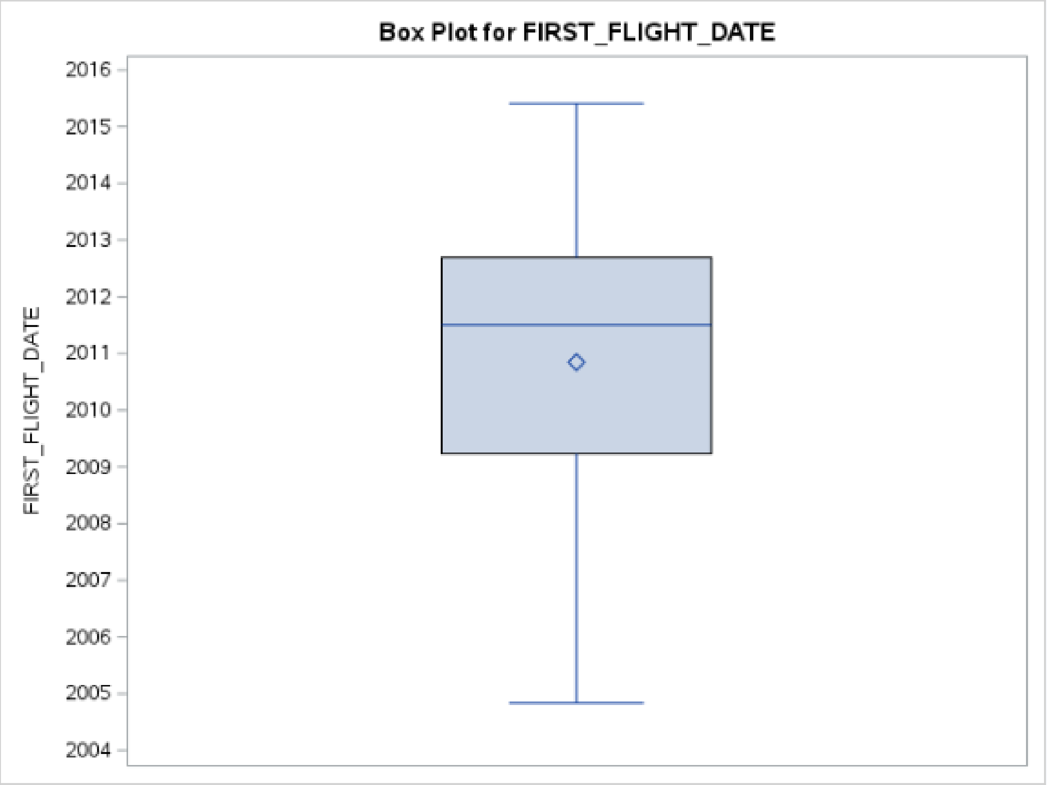


Figure 3.3.3: Updated Box Plot of FIRST_FLIGHT_DATE

In this updated box plot (Figure 3.3.3), can see that the outlier has been removed from the attribute, and it provides a more accurate representation of the dataset. The data now is showing a tighter range which is concentrated around 2009 until 2012.

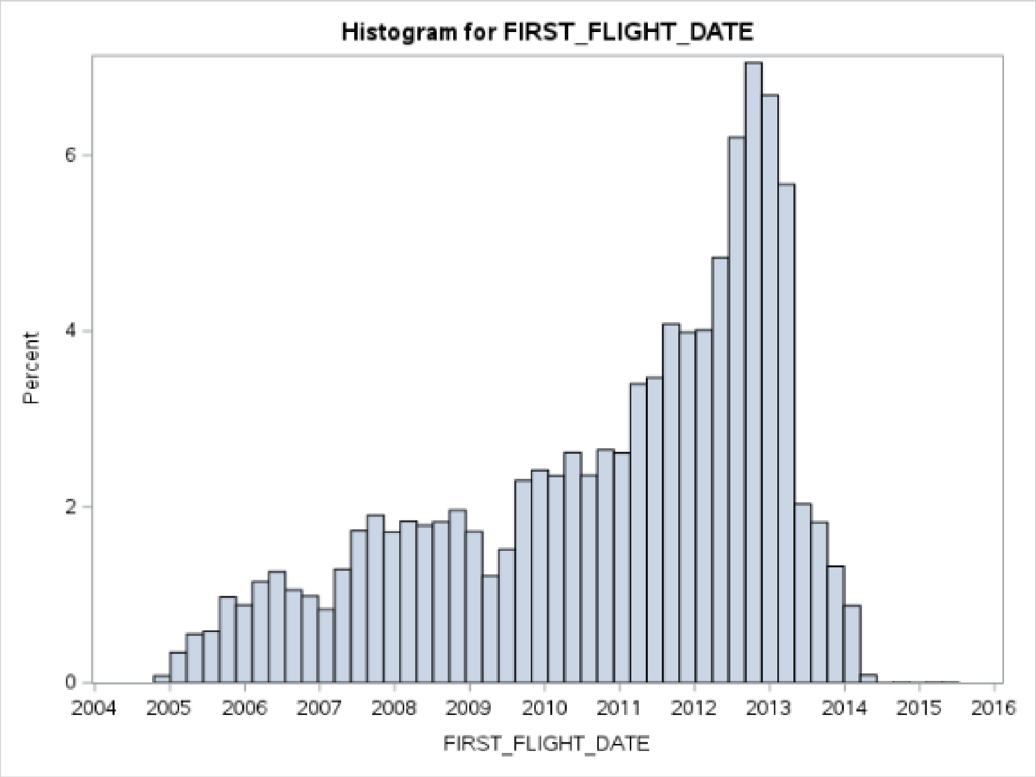
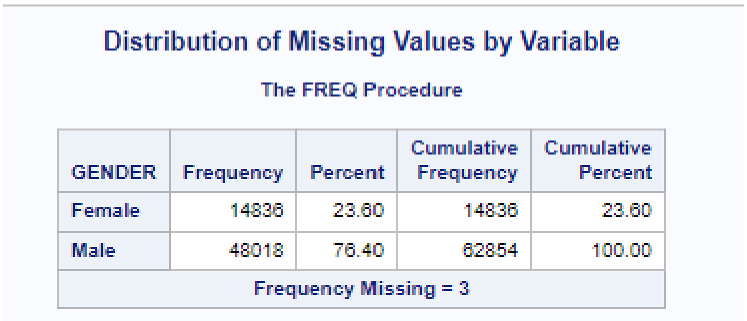


Figure 3.3.4: Updated Histogram for FIRST_FLIGHT_DATE

The updated histogram remains normally distributed with a peak around 2012 and 2013. The removal of outlier has refined the dataset and ensures that a more accurate analysis.

3.4 GENDER

Pre-Processing Method:



Distribution of Missing Values by Variable

The FREQ Procedure

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	14836	23.60	14836	23.60
Male	48018	76.40	62854	100.00
Frequency Missing = 3				

Figure 3.4.1: Check GENDER missing values

In Figure 3.4.1 it checks the distribution of missing values of GENDER. The table included Frequency (The count of occurrence by each entity), Percent (The percentage of the dataset that the entity represent), Cumulative Frequency (The running total frequency) and Cumulative Percent (The running total percentage).

Cleaning Method: Check for Missing Values

Based on the table, can clearly see that there are total of 3 missing values in GENDER attribute. Depending on the analysis requirements and the significance of the missing data, a decision will be made to either impute the missing values with a suitable method or exclude the rows with missing values.

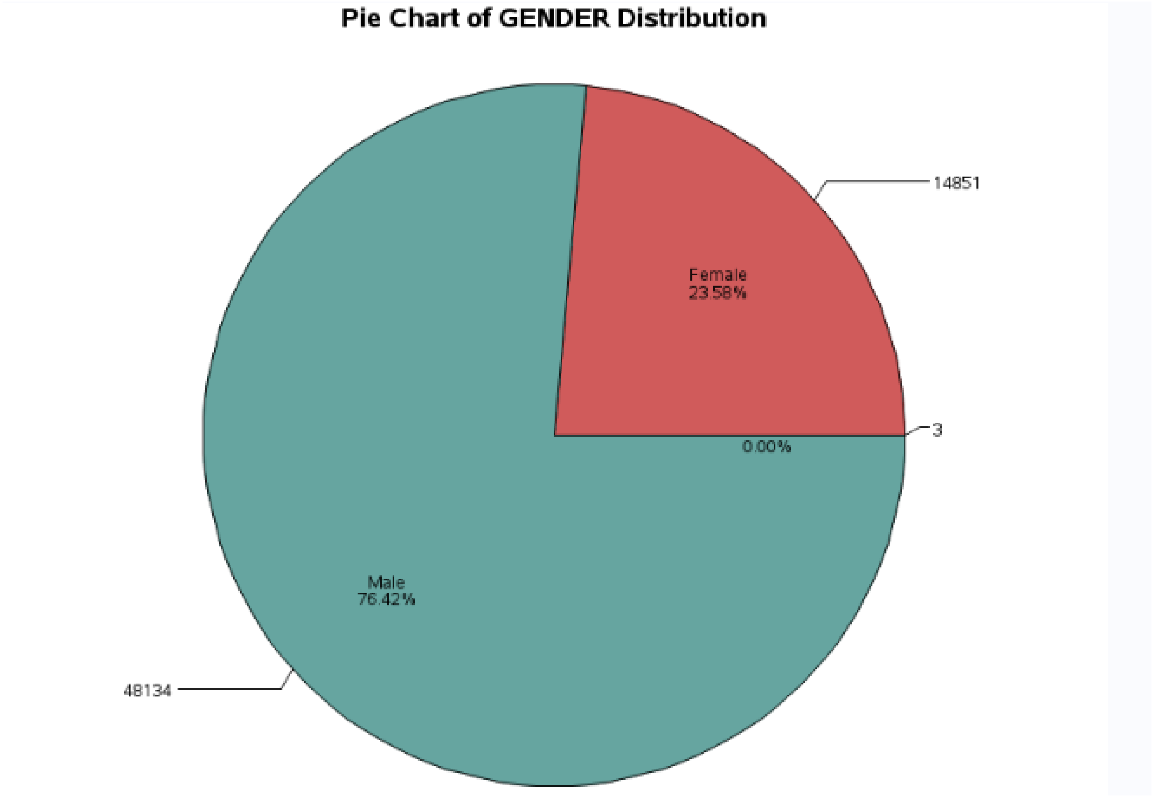


Figure 3.4.2: Pie Chart of GENDER

Frequency Analysis: Create pie chart to show the distribution of Gender.
In Figure 3.4.2, can clearly see that the Male has 48,134 of people and it has occupied 76.42%. Besides Female has a total of 14,851 of people and it occupied 23.58% in the whole column. Other than that, there are 3 missing values, but doesn't cost any percentage in the pie chart.
By having this pie chart, is to provide a visual summary of the gender distribution, to make it easier to understand the proportion of each gender within the dataset. This can help to identify any significant imbalance in the gender distribution, informing decision on whether stratified sampling might be necessary in subsequent analyses, and it will also provide a clear immediate understanding of the dataset's composition.

3.5 FFP_TIER

Before Pre-Processing:

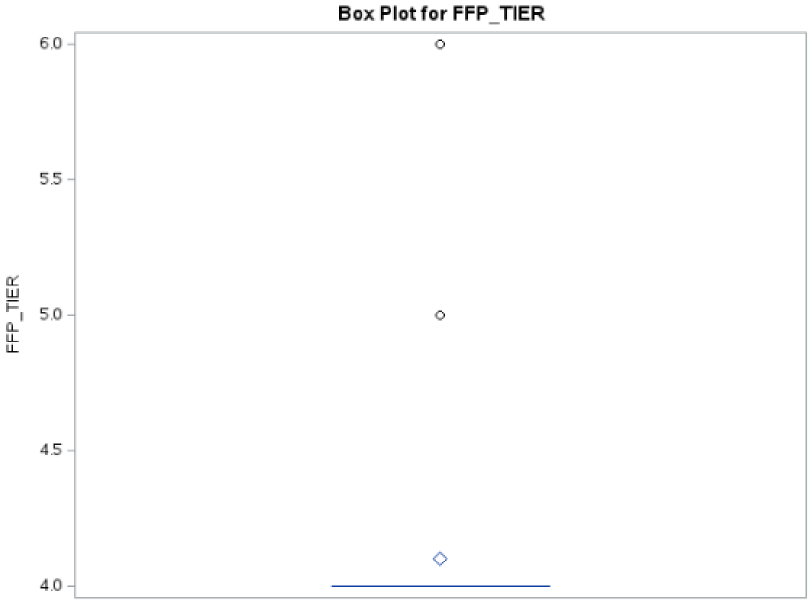


Figure 3.5.1: Box Plot of FFP_TIER

In Figure 3.5.1 it shows the box plot of the distribution of the variable or FFP_TIER. Most of the data points are at FFP_TIER 4, with a few outliers at 5 and 6.

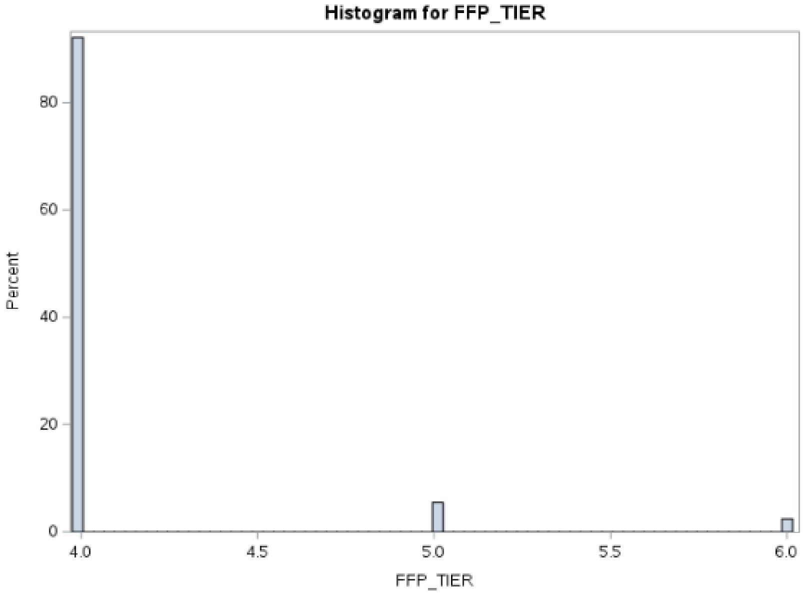


Figure 3.5.2: Histogram of FFP_TIER

In this histogram, it indicates that the majority of the FFP_TIER values are 4. However, there are some values is at 5 and 6, which are considered as outliers.

Pre-Processing Method:

Cleaning Method: General Cleaning Process

Due to make the data more useful and appropriate, in general cleaning we removed the rows that have more than 2 missing values, and overall, the missing data were checked and handled.

After General Cleaning:

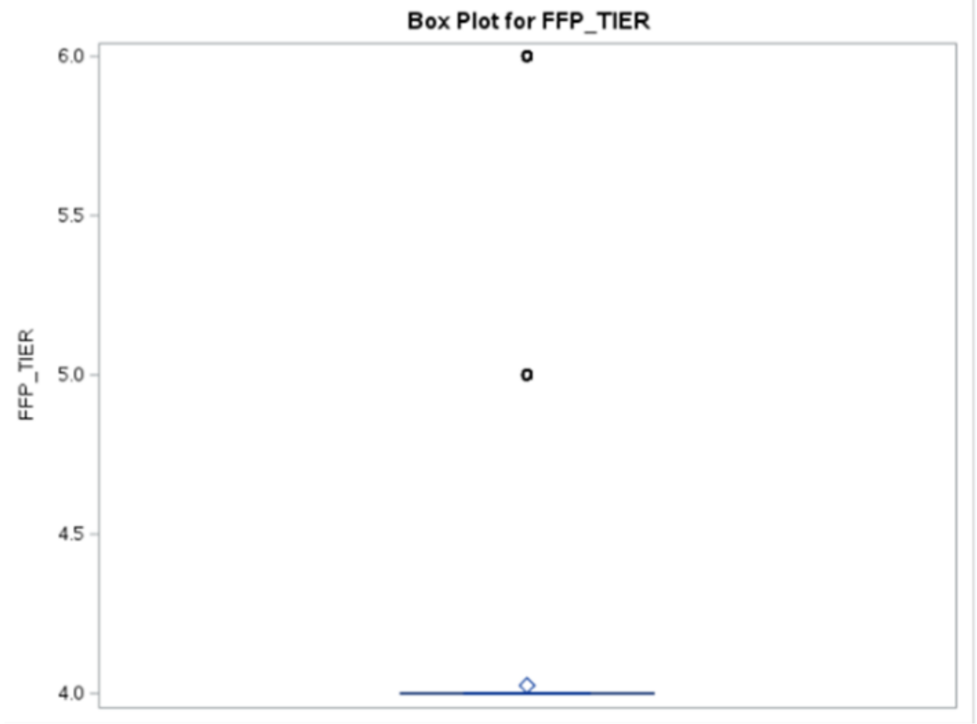


Figure 3.5.3: Updated Box Plot of FFP_TIER

In Figure 3.5.3, the updated box plot reflects the dataset after general cleaning process, showing a similar concentration at FFP_TIER 4 with minimal changes.

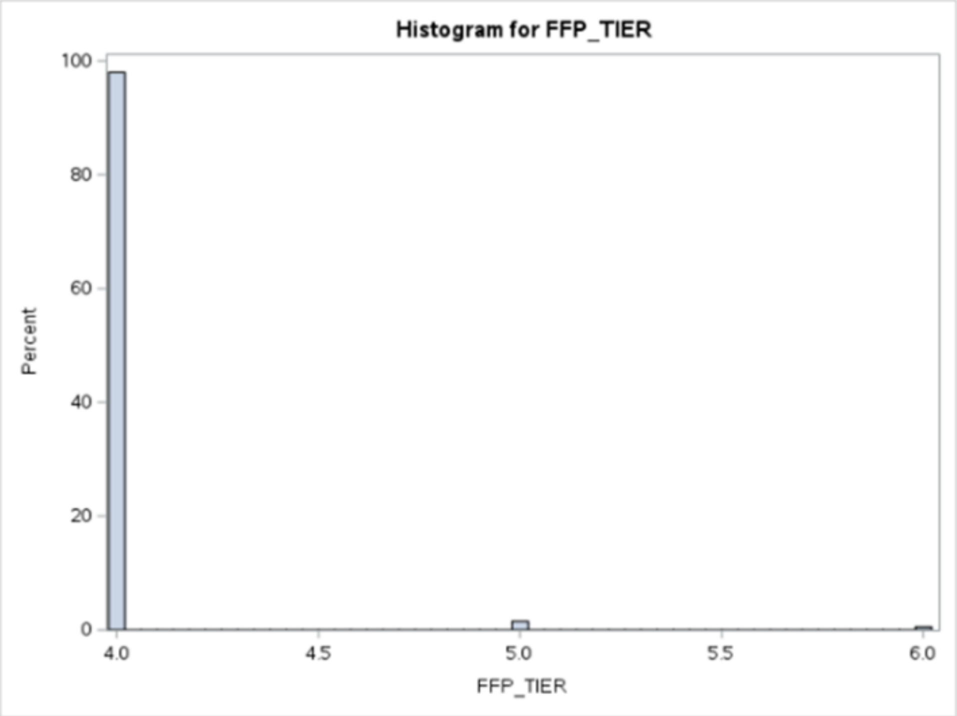
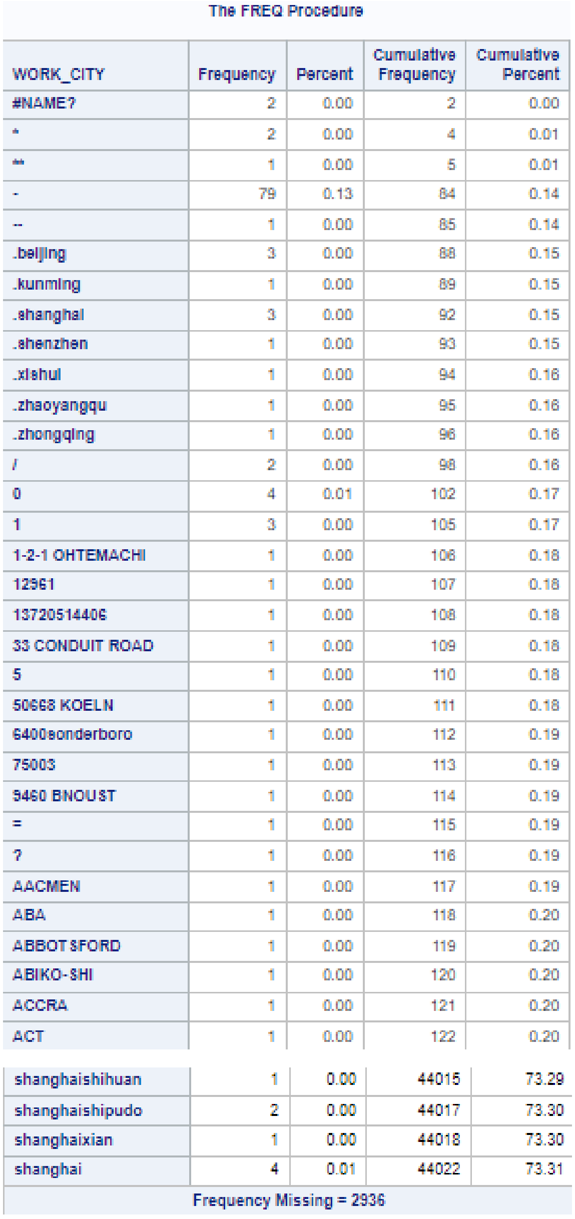


Figure 3.5.4: Updated Histogram of FFP_TIER

The updated histogram shows the distribution after general cleaning, with most of the data points still concentrated in 4.

3.6 WORK_CITY

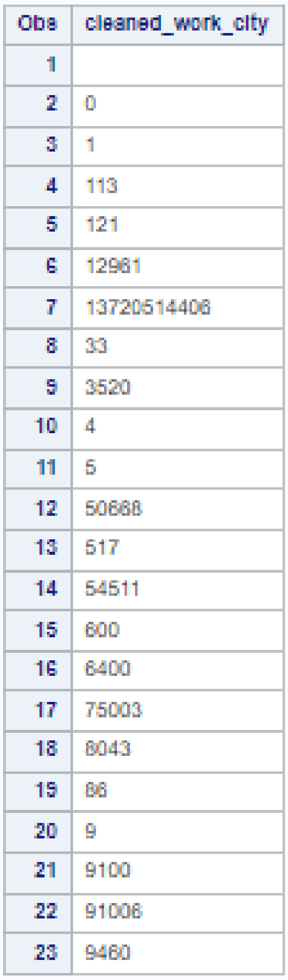
Before Pre-Processing:



The FREQ Procedure

WORK_CITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#NAME?	2	0.00	2	0.00
*	2	0.00	4	0.01
**	1	0.00	5	0.01
-	79	0.13	84	0.14
--	1	0.00	85	0.14
.beijing	3	0.00	88	0.15
.kunming	1	0.00	89	0.15
.shanghai	3	0.00	92	0.15
.shenzhen	1	0.00	93	0.15
.xishui	1	0.00	94	0.16
.zhaoyangqu	1	0.00	95	0.16
.zhongqing	1	0.00	96	0.16
/	2	0.00	98	0.16
0	4	0.01	102	0.17
1	3	0.00	105	0.17
1-2-1 OHTEMACHI	1	0.00	106	0.18
12961	1	0.00	107	0.18
13720514406	1	0.00	108	0.18
33 CONDUIT ROAD	1	0.00	109	0.18
5	1	0.00	110	0.18
50668 KOELN	1	0.00	111	0.18
6400sonderboro	1	0.00	112	0.19
75003	1	0.00	113	0.19
9460 BNOUST	1	0.00	114	0.19
=	1	0.00	115	0.19
?	1	0.00	116	0.19
AACMEN	1	0.00	117	0.19
ABA	1	0.00	118	0.20
ABBOTSFORD	1	0.00	119	0.20
ABIKO-SHI	1	0.00	120	0.20
ACCRA	1	0.00	121	0.20
ACT	1	0.00	122	0.20
shanghaishihuan	1	0.00	44015	73.29
shanghaishipudo	2	0.00	44017	73.30
shanghaixian	1	0.00	44018	73.30
shanghai	4	0.01	44022	73.31

Frequency Missing = 2936



Obs	cleaned_work_city
1	
2	0
3	1
4	113
5	121
6	12961
7	13720514406
8	33
9	3520
10	4
11	5
12	50668
13	517
14	54511
15	600
16	6400
17	75003
18	8043
19	86
20	9
21	9100
22	91006
23	9460

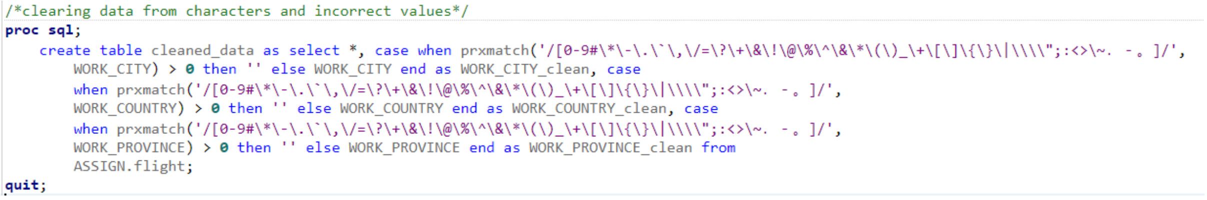
Figure 3.6.1: The Frequency Procedure of WORK_CITY

Figure 3.6.2: A portion of non-relevant data in WORK_CITY

Based on this Figure 3.6.1 and Figure 3.6.2, the frequency distribution of WORK_CITY shows numerous non-relevant entries such as symbols and integers that didn't show any city names. These invalid entries need to be cleaned to ensure the data accurately represents city names.

Pre-Processing Method:

Cleaning Method: Remove all invalid characters.

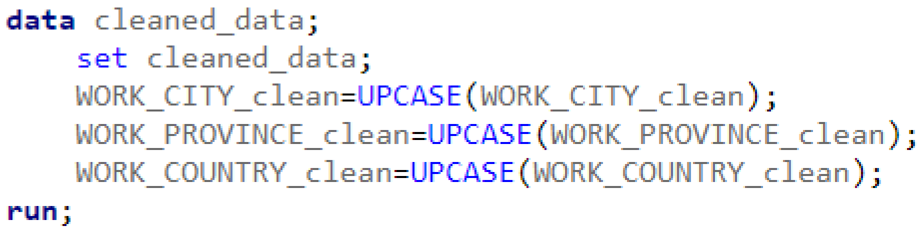


```
/*clearing data from characters and incorrect values*/
proc sql;
    create table cleaned_data as select *, case when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_CITY) > 0 then '' else WORK_CITY end as WORK_CITY_clean, case
        when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_COUNTRY) > 0 then '' else WORK_COUNTRY end as WORK_COUNTRY_clean, case
        when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_PROVINCE) > 0 then '' else WORK_PROVINCE end as WORK_PROVINCE_clean from
        ASSIGN.flight;
quit;
```

Figure 3.6.3: Code of Removing Invalid Characters

The initial step in cleaning WORK_CITY non-relevant data is removing all invalid characters, such as symbols and integer from the WORK_CITY attribute. By cleaning it, we use ‘proc sql’ to demonstrate the process of creating a new cleaned table called WORK_CITY_clean by removing all invalid characters using regular expressions.

Cleaning Method: Uppercase cleaned variables

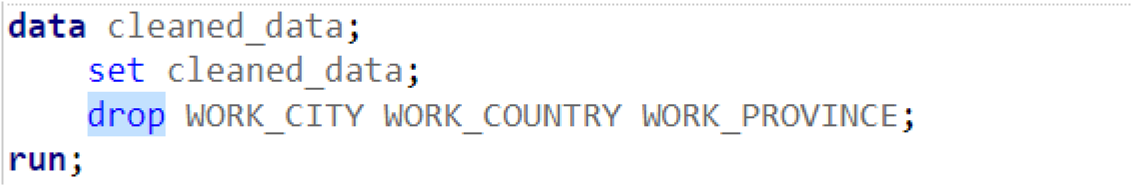


```
data cleaned_data;
    set cleaned_data;
    WORK_CITY_clean=UPCASE(WORK_CITY_clean);
    WORK_PROVINCE_clean=UPCASE(WORK_PROVINCE_clean);
    WORK_COUNTRY_clean=UPCASE(WORK_COUNTRY_clean);
run;
```

Figure 3.6.4: Codes of Transforming data to Uppercase.

To maintain consistency and make the data more united, all entire in the cleaned WORK_CITY_clean column will be converted to uppercase.

Cleaning Method: Drop Original Variable

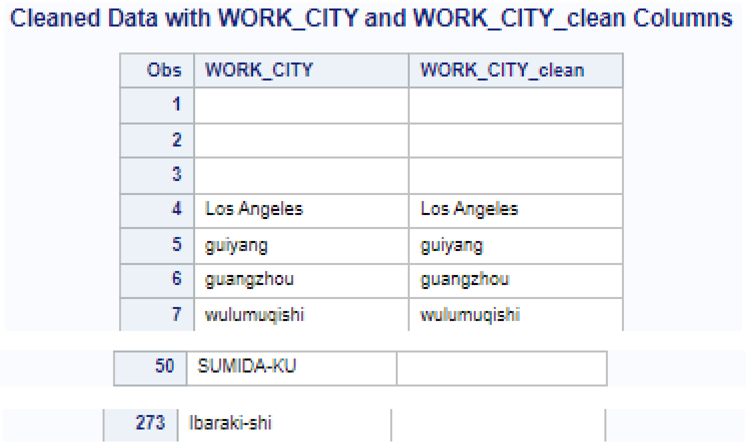


```
data cleaned_data;
    set cleaned_data;
    drop WORK_CITY WORK_COUNTRY WORK_PROVINCE;
run;
```

Figure 3.6.5: Drop the original column from table

After cleaning the data in “WORK_CITY_CLEAN”, the “WORK_CITY” variable will be dropped and replaced it with the clean column, which was “WORK_CITY_CLEAN” column.

After Pre-Processing:

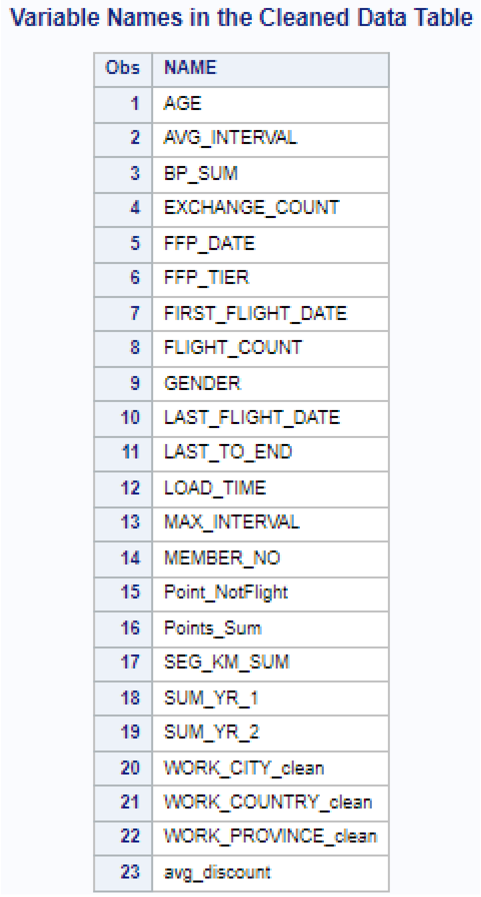


Cleaned Data with WORK_CITY and WORK_CITY_clean Columns

Obs	WORK_CITY	WORK_CITY_clean
1		
2		
3		
4	Los Angeles	Los Angeles
5	guiyang	guiyang
6	guangzhou	guangzhou
7	wulumuqishi	wulumuqishi
50	SUMIDA-KU	
273	Ibaraki-shi	

Figure 3.6.6: Comparison of before and after cleaning

In Figure 3.6.6 shows that the comparison of WORK_CITY and WORK_CITY_clean. The clean table shows that the column with invalid characters and integers had been removed. Other than that, the invalid city names were replaced with missing values too, this is to ensure the data integrity.

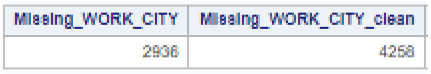


Variable Names in the Cleaned Data Table

Obs	NAME
1	AGE
2	AVG_INTERVAL
3	BP_SUM
4	EXCHANGE_COUNT
5	FFP_DATE
6	FFP_TIER
7	FIRST_FLIGHT_DATE
8	FLIGHT_COUNT
9	GENDER
10	LAST_FLIGHT_DATE
11	LAST_TO_END
12	LOAD_TIME
13	MAX_INTERVAL
14	MEMBER_NO
15	Point_NotFlight
16	Points_Sum
17	SEG_KM_SUM
18	SUM_YR_1
19	SUM_YR_2
20	WORK_CITY_clean
21	WORK_COUNTRY_clean
22	WORK_PROVINCE_clean
23	avg_discount

Figure 3.6.7: The variable names after cleaning

In the Variable Names in the Cleaned Data Table, can see that the attribute name had included WORK_CITY_clean and the WORK_CITY column had been removed from the original column.



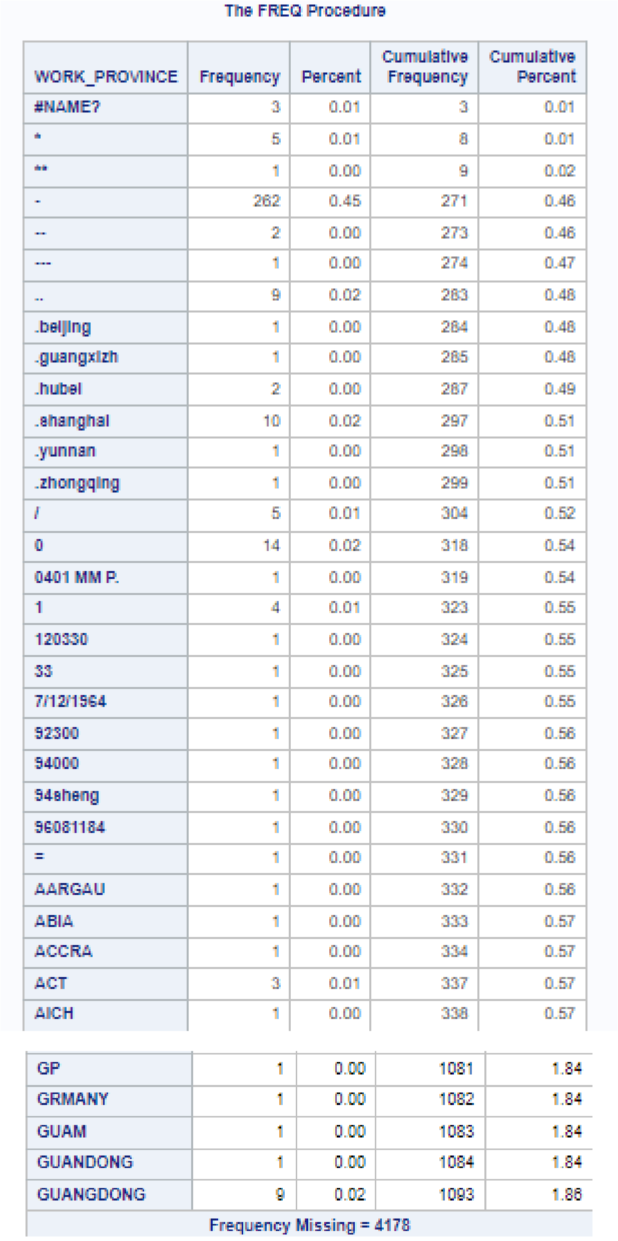
Missing_WORK_CITY	Missing_WORK_CITY_clean
2936	4258

Figure 3.6.8: The total number of missing values after cleaned

In conclusion, the frequency of missing values increased to 4258 after the cleaning process, due to the fact that the invalid entries had been removed.

3.7 WORK_PROVINCE

Before Pre-Processing:



The FREQ Procedure

WORK_PROVINCE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
#NAME?	3	0.01	3	0.01
*	5	0.01	8	0.01
**	1	0.00	9	0.02
-	262	0.45	271	0.46
--	2	0.00	273	0.46
---	1	0.00	274	0.47
..	9	0.02	283	0.48
.beijing	1	0.00	284	0.48
.guangxizh	1	0.00	285	0.48
.hubei	2	0.00	287	0.49
.shanghai	10	0.02	297	0.51
.yunnan	1	0.00	298	0.51
.zhongqing	1	0.00	299	0.51
/	5	0.01	304	0.52
0	14	0.02	318	0.54
0401 MM P.	1	0.00	319	0.54
1	4	0.01	323	0.55
120330	1	0.00	324	0.55
33	1	0.00	325	0.55
7/12/1964	1	0.00	326	0.55
92300	1	0.00	327	0.56
94000	1	0.00	328	0.56
94sheng	1	0.00	329	0.56
96081184	1	0.00	330	0.56
=	1	0.00	331	0.56
AARGAU	1	0.00	332	0.56
ABIA	1	0.00	333	0.57
ACCRA	1	0.00	334	0.57
ACT	3	0.01	337	0.57
AICH	1	0.00	338	0.57
GP	1	0.00	1081	1.84
GRMANY	1	0.00	1082	1.84
GUAM	1	0.00	1083	1.84
GUANDONG	1	0.00	1084	1.84
GUANGDONG	9	0.02	1093	1.86

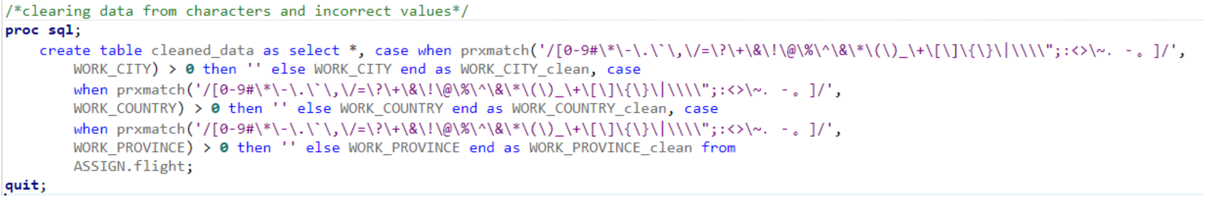
Frequency Missing = 4178

Figure 3.7.1: The Frequency Procedure of WORK_PROVINCE

For WORK_PROVINCE data, it is similar to WORK_CITY concept. Based on the Figure 3.7.1, the data included lots of non-relevant data including integers, date and symbols. These invalid entries need to be addressed to ensure the data accurately represents valid province names.

Pre-Processing Method:

Cleaning Method: Remove all invalid characters.

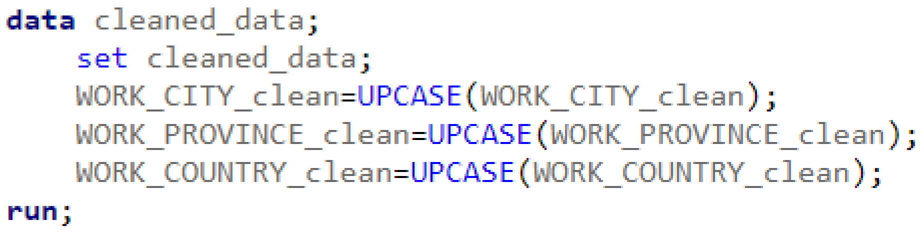


```
/*clearing data from characters and incorrect values*/
proc sql;
    create table cleaned_data as select *, case when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_CITY) > 0 then '' else WORK_CITY end as WORK_CITY_clean, case
        when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_COUNTRY) > 0 then '' else WORK_COUNTRY end as WORK_COUNTRY_clean, case
        when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_PROVINCE) > 0 then '' else WORK_PROVINCE end as WORK_PROVINCE_clean from
        ASSIGN.flight;
quit;
```

Figure 3.7.2: Codes that remove invalid characters

To clean the data, firstly all the characters that are not supposed to appear in WORK_PROVINCE attribute needs to be removed. By removing the data, we use a similar approach as WORK_CITY, and create a new attribute named WORK_PROVINCE_clean to store the cleaned data.

Cleaning Method: Uppercase cleaned variables.

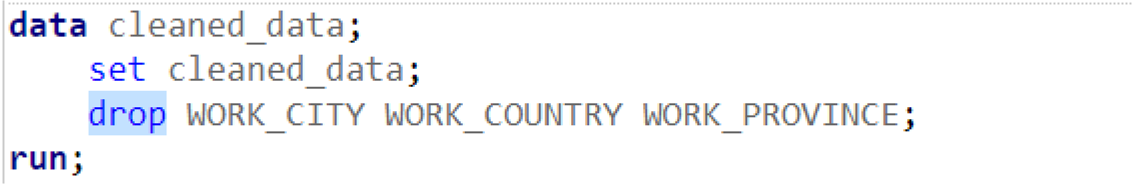


```
data cleaned_data;
    set cleaned_data;
    WORK_CITY_clean=UPCASE(WORK_CITY_clean);
    WORK_PROVINCE_clean=UPCASE(WORK_PROVINCE_clean);
    WORK_COUNTRY_clean=UPCASE(WORK_COUNTRY_clean);
run;
```

Figure 3.7.3: Codes to transform data to Uppercase

To standardize the data, all data that exists in the WORK_PROVINCE_clean column will be converted to uppercase. This is to ensure the stable and simplified comparison between entries.

Cleaning Method: Drop Original Variable.

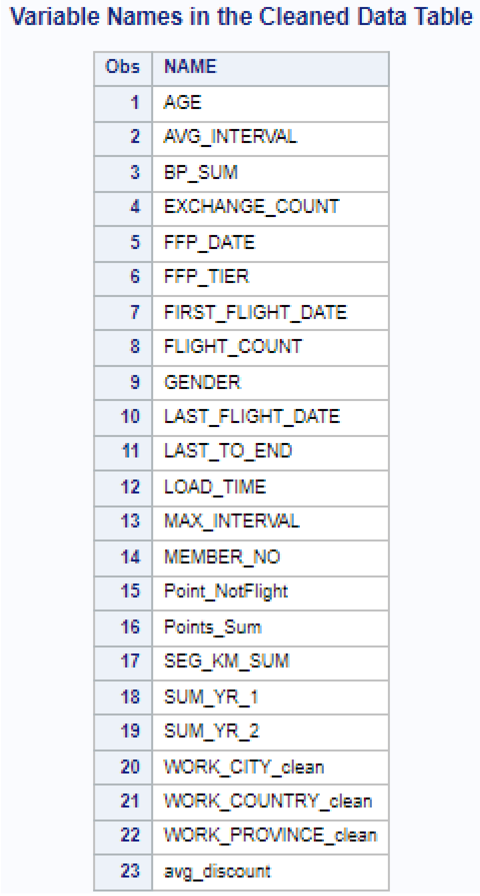


```
data cleaned_data;
    set cleaned_data;
    drop WORK_CITY WORK_COUNTRY WORK_PROVINCE;
run;
```

Figure 3.7.4: Drop the original column from table

After cleaning the data in "WORK_PROVINCE_CLEAN", the original data "WORK_PROVINCE" was dropped, and replaced with the cleaned "WORK_PROVINCE_CLEAN" column.

After Pre-Processing:

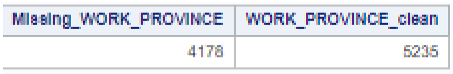


Variable Names in the Cleaned Data Table

Obs	NAME
1	AGE
2	AVG_INTERVAL
3	BP_SUM
4	EXCHANGE_COUNT
5	FFP_DATE
6	FFP_TIER
7	FIRST_FLIGHT_DATE
8	FLIGHT_COUNT
9	GENDER
10	LAST_FLIGHT_DATE
11	LAST_TO_END
12	LOAD_TIME
13	MAX_INTERVAL
14	MEMBER_NO
15	Point_NotFlight
16	Points_Sum
17	SEG_KM_SUM
18	SUM_YR_1
19	SUM_YR_2
20	WORK_CITY_clean
21	WORK_COUNTRY_clean
22	WORK_PROVINCE_clean
23	avg_discount

Figure 3.7.5: The variable names after cleaning

In Figure 3.7.5, shows that the Variable Names in the Cleaned Data Table had dropped the original column, which was WORK_PROVINCE and been replaced by WORK_PROVINCE_clean.



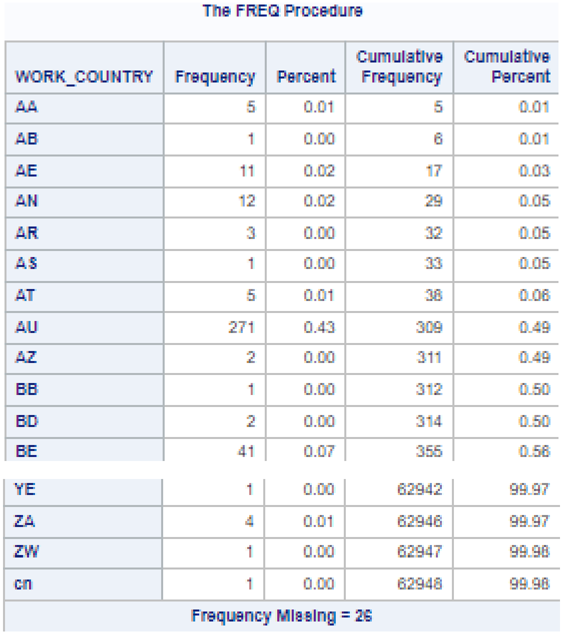
Missing_WORK_PROVINCE	WORK_PROVINCE_clean
4178	5235

Figure 3.7.6: The total number of missing values after cleaned

The frequency of missing values increased to 5235 after cleaning process, indicating the removal of invalid entries.

3.8 WORK_COUNTRY

Before Pre-Processing:



The FREQ Procedure

WORK_COUNTRY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AA	5	0.01	5	0.01
AB	1	0.00	6	0.01
AE	11	0.02	17	0.03
AN	12	0.02	29	0.05
AR	3	0.00	32	0.05
AS	1	0.00	33	0.05
AT	5	0.01	38	0.06
AU	271	0.43	309	0.49
AZ	2	0.00	311	0.49
BB	1	0.00	312	0.50
BD	2	0.00	314	0.50
BE	41	0.07	355	0.56
YE	1	0.00	62942	99.97
ZA	4	0.01	62946	99.97
ZW	1	0.00	62947	99.98
cn	1	0.00	62948	99.98

Frequency Missing = 26

Figure 3.8.1: The Frequency Procedure of WORK_COUNTRYY

For WORK_COUNTRY, the attribute contains inconsistent entries which was lowercase letter. In Figure 3.8.1, the tables show that there are total of 26 missing values in the WORK_COUNTRY table. Additionally, the table contains inconsistent entries, specifically lowercase letters, which can cause inaccuracies in data analysis.

Based on the Figure 3.8.1, you can see there are 4 columns are analysing the WORK_COUNTRY table:

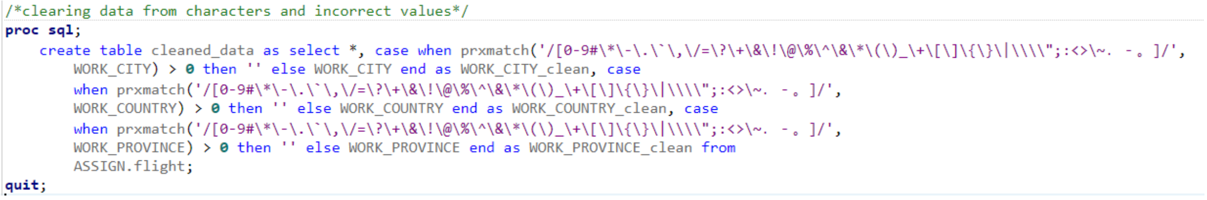
1 Frequency: This column shows the number of occurrences of each unique value that existed in WORK_COUNTRY.

2 Percent: This column shows the percentage of each unique value from WORK_COUNTRY that relative to the total.

3 Cumulative Frequency and Cumulative Percent: Two of these columns are similar, it shows the running total and the cumulative percentage in a separate way.

Pre-Processing Method:

Cleaning Method: Remove all invalid characters.

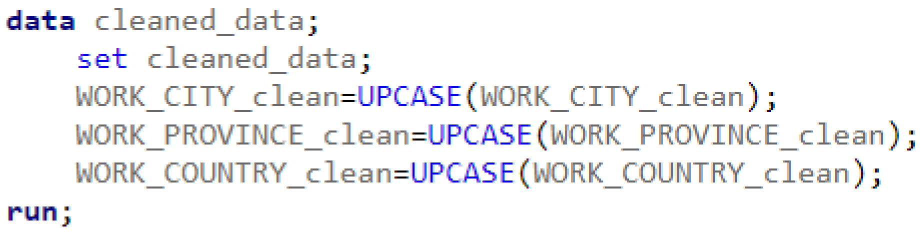


```
/*clearing data from characters and incorrect values*/
proc sql;
    create table cleaned_data as select *, case when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_CITY) > 0 then '' else WORK_CITY end as WORK_CITY_clean, case
        when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_COUNTRY) > 0 then '' else WORK_COUNTRY end as WORK_COUNTRY_clean, case
        when prxmatch('/[0-9#\*\-\.\`\,\/=\?\+\&\!\@\%\^\&\*\(\)_\+\[\]\{\}\|\\\\";:<>\~. -。]/',
        WORK_PROVINCE) > 0 then '' else WORK_PROVINCE end as WORK_PROVINCE_clean from
        ASSIGN.flight;
quit;
```

Figure 3.8.2: Codes that remove invalid characters

By having this step is important, the main purpose is to remove all irrelevant or non-logical data from the attribute. This is to ensure that only valid and meaningful entries are retained.

Cleaning Method: Uppercase cleaned variables

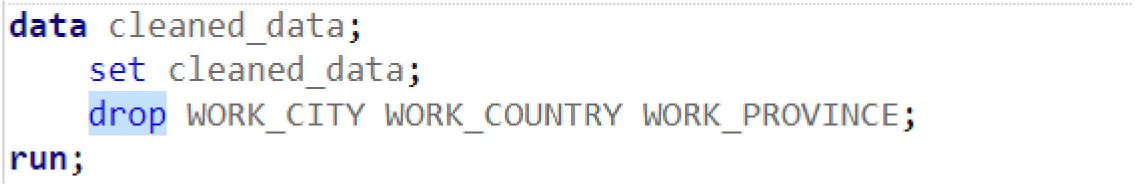


```
data cleaned_data;
    set cleaned_data;
    WORK_CITY_clean=UPCASE(WORK_CITY_clean);
    WORK_PROVINCE_clean=UPCASE(WORK_PROVINCE_clean);
    WORK_COUNTRY_clean=UPCASE(WORK_COUNTRY_clean);
run;
```

Figure 3.8.3: Codes to transform data to Uppercase

Secondly, the next step of cleaning is to convert all of the WORK_COUNTRY_clean column to uppercase. This standardizes the data, making it uniform and simplifying comparisons between entities.

Cleaning Method: Drop Original Variable

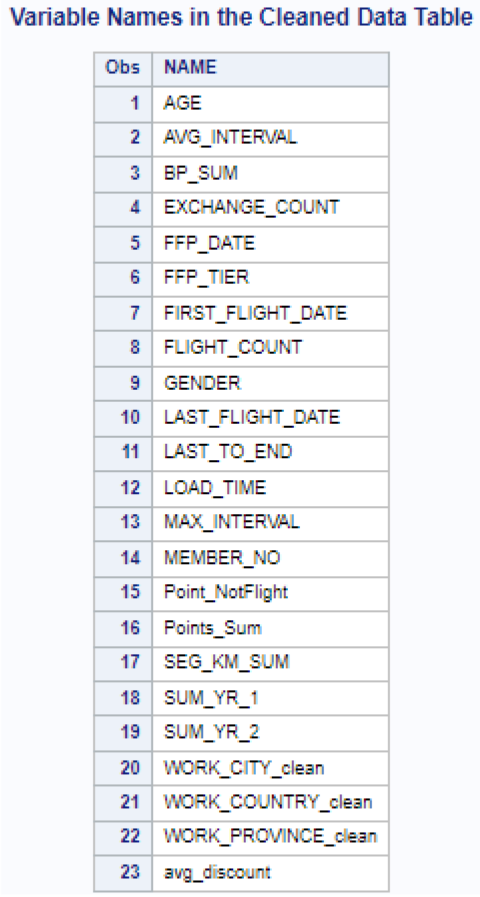


```
data cleaned_data;
    set cleaned_data;
    drop WORK_CITY WORK_COUNTRY WORK_PROVINCE;
run;
```

Figure 3.8.4: Drop the original column from table

Thirdly, to avoid duplicating data, the original data from WORK_COUNTRY column will be removed from the table and the cleaned version, which was "WORK_COUNTRY_clean" will be replacing it.

After Pre-Processing:



Variable Names in the Cleaned Data Table

Obs	NAME
1	AGE
2	AVG_INTERVAL
3	BP_SUM
4	EXCHANGE_COUNT
5	FFP_DATE
6	FFP_TIER
7	FIRST_FLIGHT_DATE
8	FLIGHT_COUNT
9	GENDER
10	LAST_FLIGHT_DATE
11	LAST_TO_END
12	LOAD_TIME
13	MAX_INTERVAL
14	MEMBER_NO
15	Point_NotFlight
16	Points_Sum
17	SEG_KM_SUM
18	SUM_YR_1
19	SUM_YR_2
20	WORK_CITY_clean
21	WORK_COUNTRY_clean
22	WORK_PROVINCE_clean
23	avg_discount

Figure: 3.8.5: Cleaned Date Table

In the Figure 3.8.5, can see that the original attribute had been removed, and the cleaned data has shown in the table. Now the data is free from lowercase data, ensuring more accurate and reliable analysis in the future.

3.9 AGE

Before Pre-Processing:

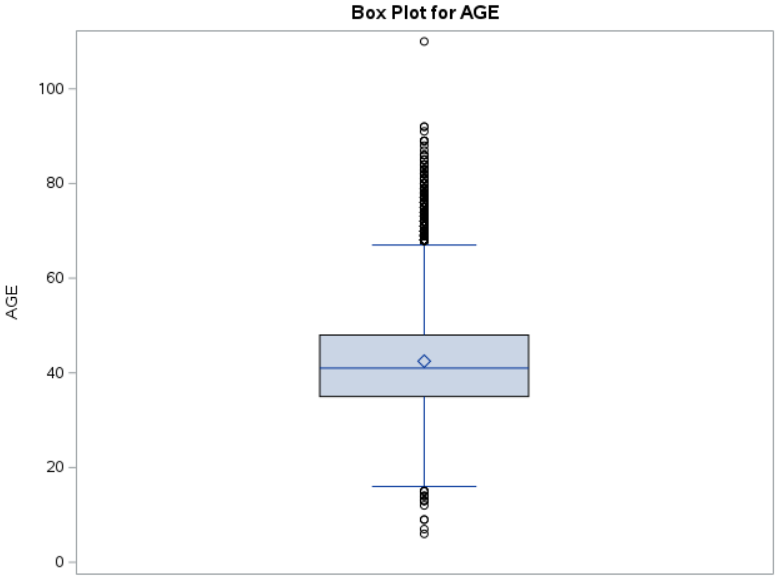


Figure 3.9.1: Box Plot of AGE

The box plot of AGE (Figure 3.9.1) indicates the presence of numerous outliers, as represented by the points outside the whiskers. These outliers suggest that there are individuals in the dataset with ages that are significantly higher or lower than the majority.

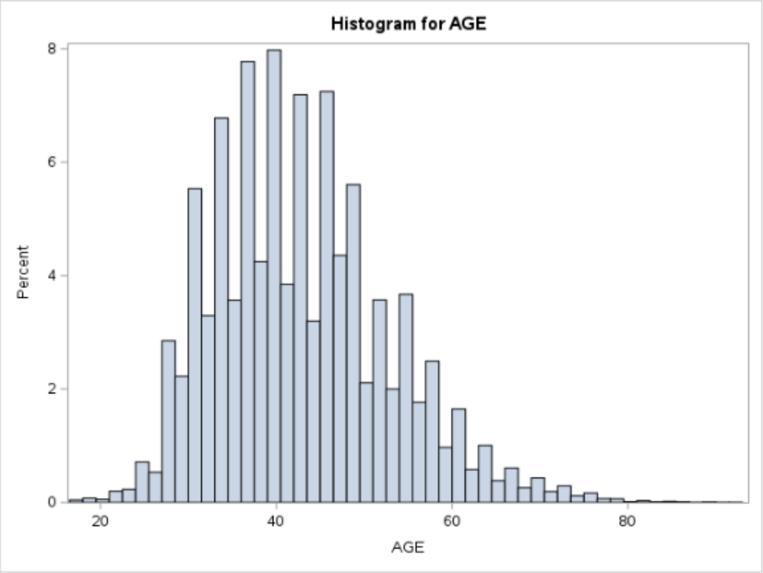
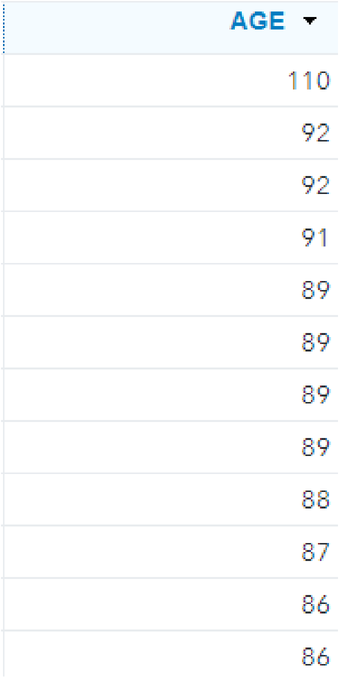


Figure 3.9.2: Histogram of AGE

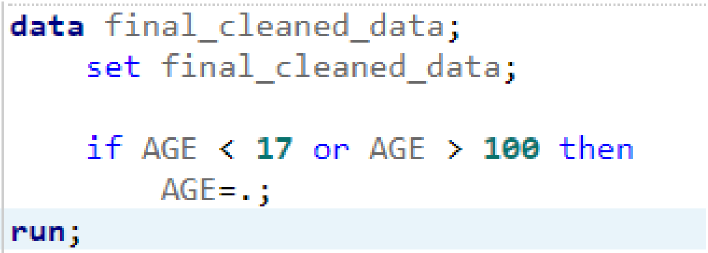
In Figure 3.9.2, the histogram of AGE shows the distribution of ages within the dataset. The histogram performs normally distributed with a peak around 40s, which is also make sense that

40s usually will work, and may take a plane for working purpose. However, if observed carefully, there are some noticeable tails at both of the ends, this r4esults some extreme values appear in the column.

Pre-processing method:



AGE
110
92
92
91
89
89
89
89
88
87
86
86



```
data final_cleaned_data;
    set final_cleaned_data;

    if AGE < 17 or AGE > 100 then
        AGE=.;
run;
```

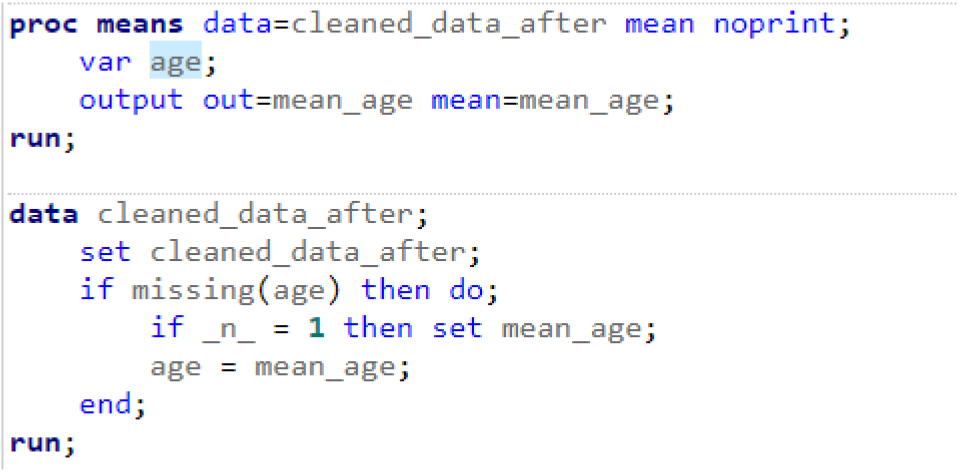
Figure 3.9.3: The outlier data that existed

Figure 3.9.4: The IQR Method for Removing Outliers

Cleaning method: IQR Method for removing Outliers.

To address the issues of outliers, the Interquartile Range (IQR) method was applied. This method helps to identify and remove data points that are significantly lower or higher than the mass of the data.

Given that that is not logical to expect individuals younger than 18 or older than 100 to be in this dataset, due to the fact that is it illogical that these ages of individual working. Therefore, the ages or the data that outside this range will be seen as an invalid data or outliers and will be removed from the attribute.



```
proc means data=cleaned_data_after mean noprint;
    var age;
    output out=mean_age mean=mean_age;
run;

data cleaned_data_after;
    set cleaned_data_after;
    if missing(age) then do;
        if _n_ = 1 then set mean_age;
        age = mean_age;
    end;
run;
```

Figure 3.9.5: Code of replacing missing values

Cleaning method: Replace missing value using average value

After cleaning the extreme outliers, the next step was to fill in the missing values in the AGE column by using mean. The mean age will be calculated and used to fill in the missing values in the column. This is to ensure that the dataset remains robust and can be used effectively for further analysis in the future.

After Pre-processing:

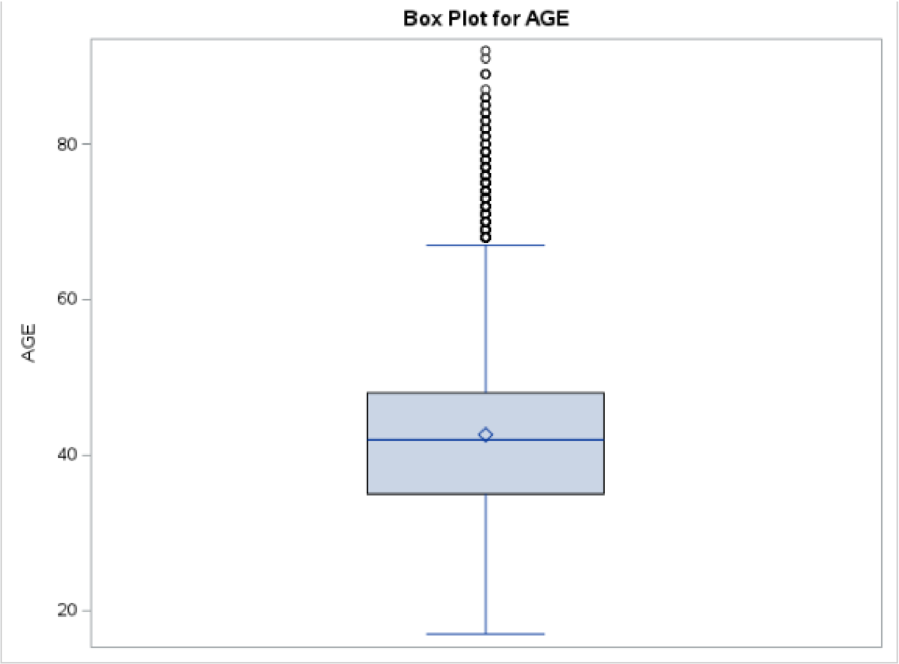


Figure 3.9.6: Updated Box Plot of AGE

After cleaning the data, the updated box plot of AGE (Figure 3.9.6) shows the reduction in the number of outliers. The extreme ages have been removed, resulting in a more accurate representation of the age distribution within the dataset.

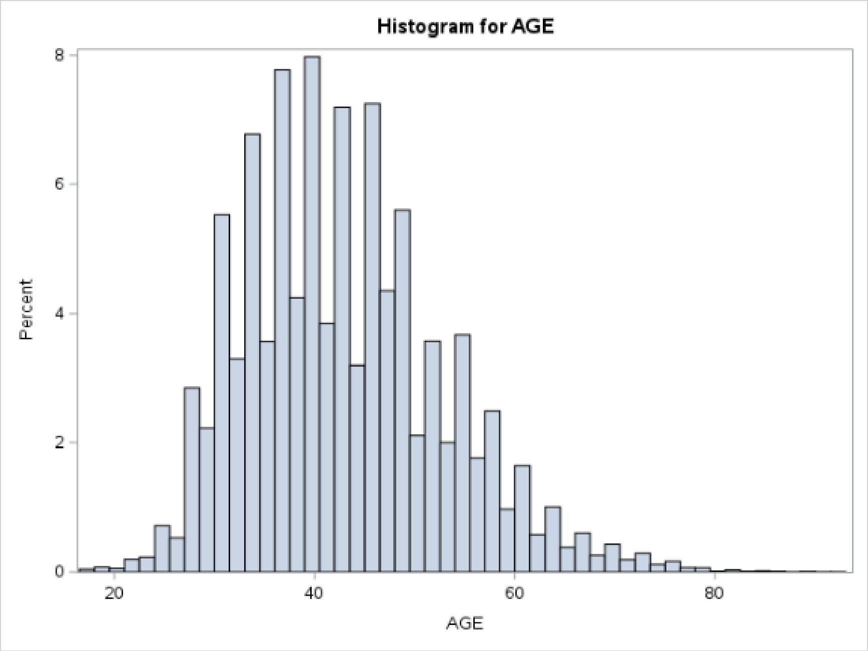


Figure 3.9.7: Updated Histogram of AGE

The updated histogram of AGE (Figurer 3.9.7) displays a more focused distribution. The previous histogram can see that the tails have been minimized, indicating that the current data is better to reflects a realistic age range.

3.10 LOAD_TIME

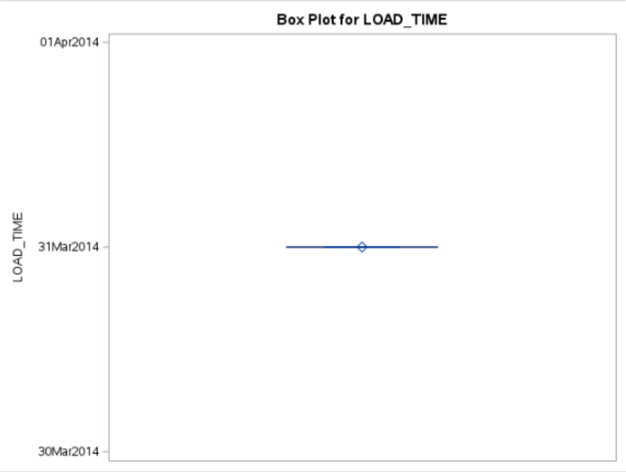


Figure 3.10.1: Box Plot of LOAD_TIME

The box plot shows that all values for LOAD_TIME fall within a narrow range, and there are not significant outliers exist in LOAD_TIME.

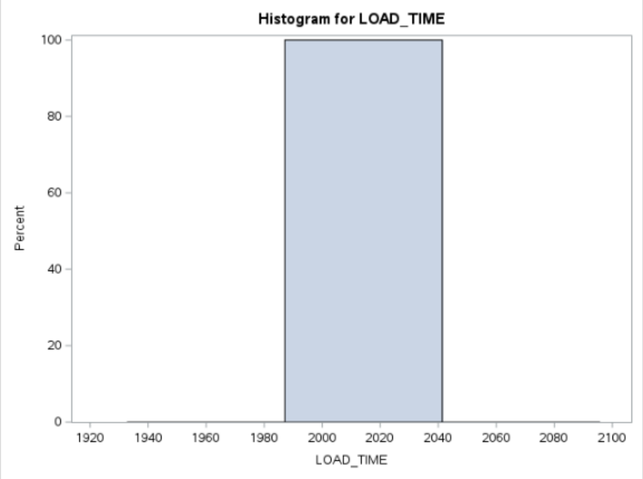


Figure 3.10.2: Histogram of LOAD_TIME

The histogram shows that LOAD_TIME values are concentrated within a specific range, indicating consistent data with no variability.

Pre-Processing Method:
Cleaning Method: Checked for missing values.
Due to this LOAD_TIME data doesn't exist any missing values and the data distribution is clean and consistent, there are no pre-processing was required for LOAD_TIME.

3.11 FLIGHT_COUNT

Before Pre-Processing:

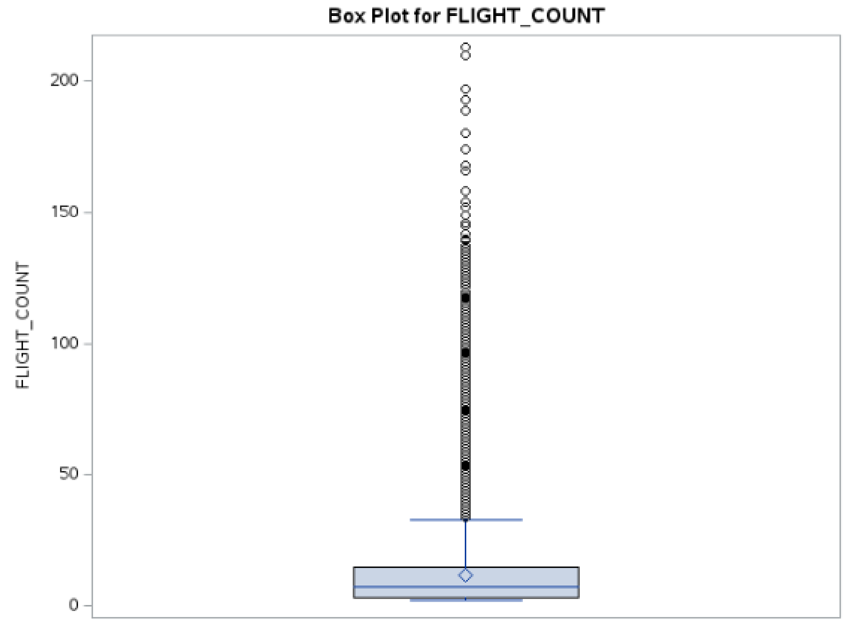


Figure 3.11.1: Box Plot of FLIGHT_COUNT

In Figure 3.11.1, the box plot of FLIGHT_COUNT contains a significant number of outliers. Outliers that shows are unusually high compared to the rest of the data. The presence of these outliers indicates that some of the value in this dataset does not fit in the general pattern of FLIGHT_COUNT.

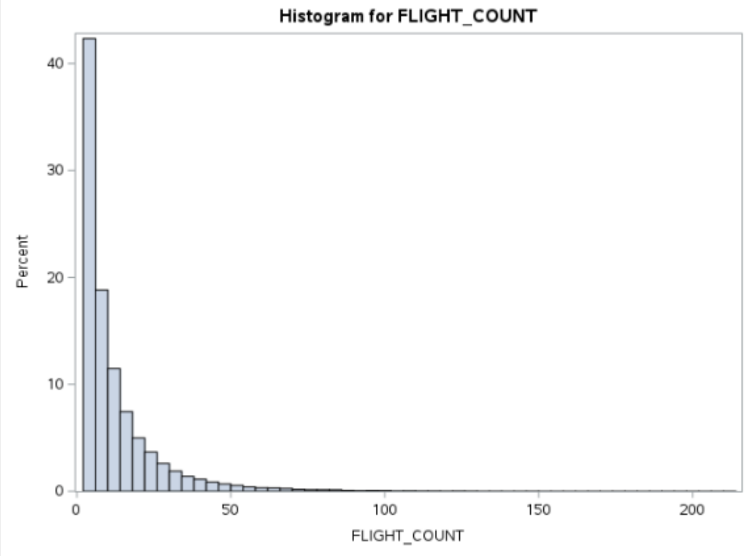


Figure 3.11.2: Histogram of FLIGHT_COUNT

In Figure 3.11.2, the histogram shows the distribution of FLIGHT_COUNT before implementing or editing by using any pre-processing method. Can clearly see that many FLIGHT_COUNT values are clustered around the lower end with a long tail extending towards the higher value. By observing this distribution can confirm that the presence of outliers, as the long tail suggests some data points are much higher than the majority value.

Pre-Processing Method:

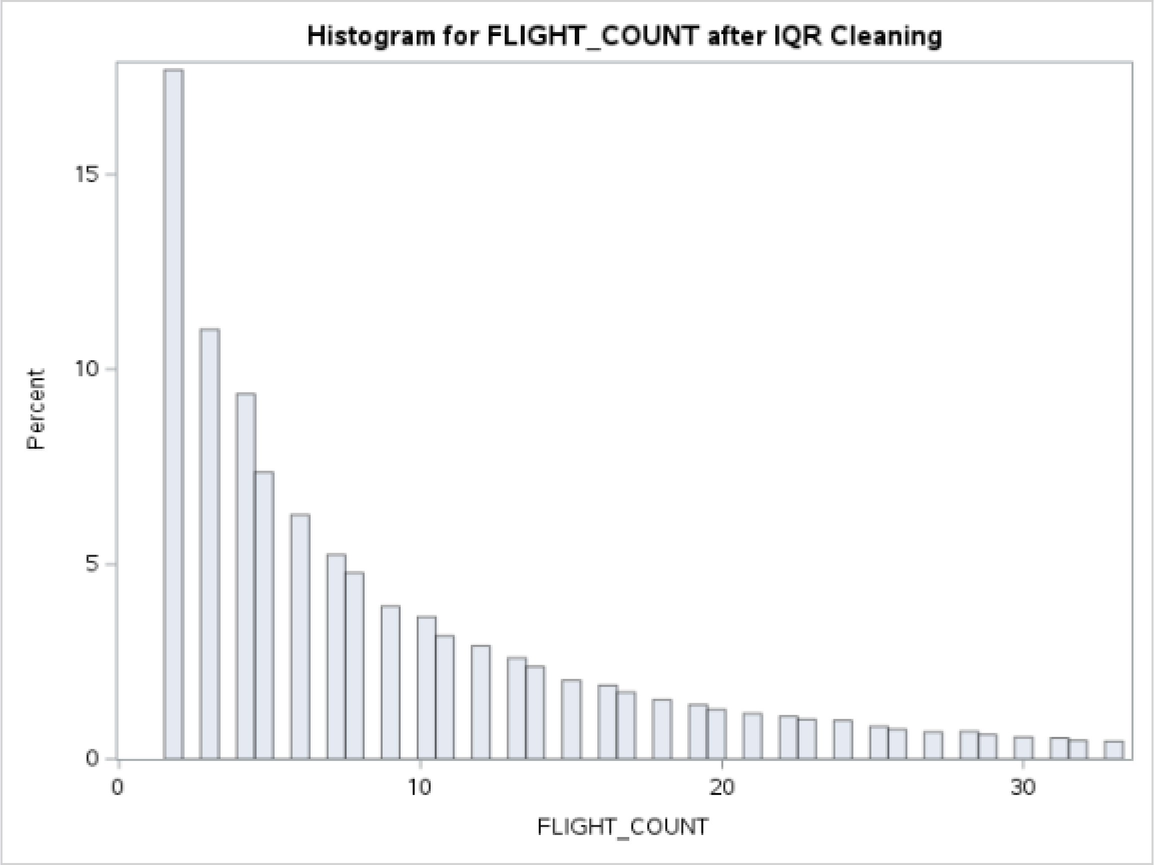


Figure 3.11.3: Histogram of FLIGHT_COUNT after IQR Cleaning

Cleaning Method: IQR Method for removing Outliers.

The Interquartile Range (IQR) method was applied in the FLIGHT_COUNT to remove the outliers from the dataset. This method involves calculating the IQR, which is the range between the first quartile (25 percent) and the third quartile (75 percentage). By removing the outliers, the data point which lies more than 1.5 times the third quartile or below the first quartile will be considered as outlier and will be removed by using IQR method.

After Pre-Processing:

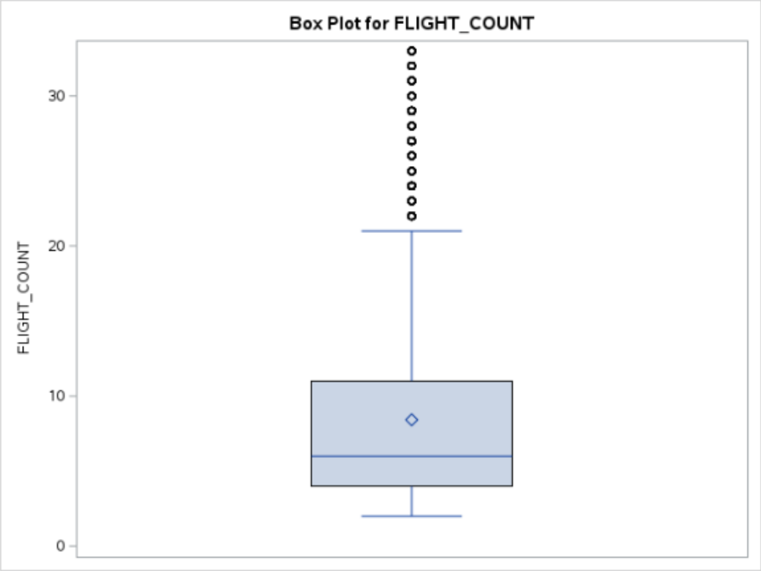


Figure 3.11.4: Box Plot of FLIGHT_COUNT

In Figure (3.11.4), is the updated box plot of FLIGHT_COUNT. After applying the IQR method, can clearly see that the outliers of FLIGHT_COUNT are lesser than before in the Figure (3.11.1). Now the data is more tightly grouped with the majority of values falling within a narrower range.

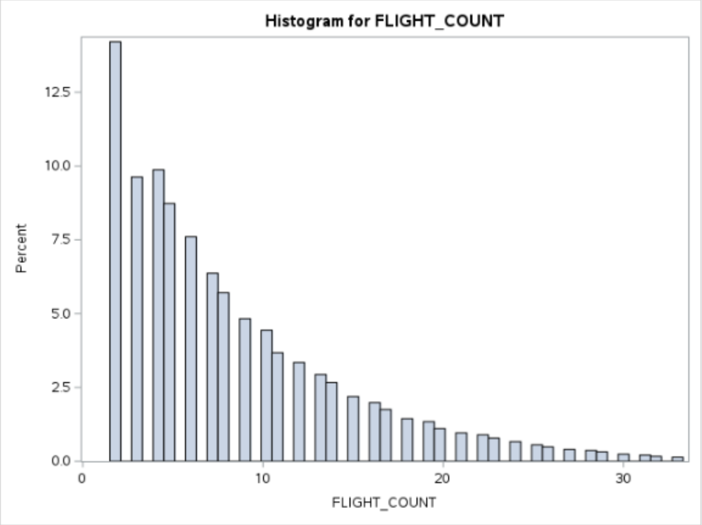


Figure 3.11.5: Histogram of FLIGHT_COUNT

After using the IQR method, the histogram of FLIGHT_COUNT shows a more stable and constant distribution. The removal of outliers has resulted in a dataset where the majority of values are closer to each other. In this cleaned dataset, the majority of values are now closer to each other.

3.12 BP_SUM

Before Pre-Processing:

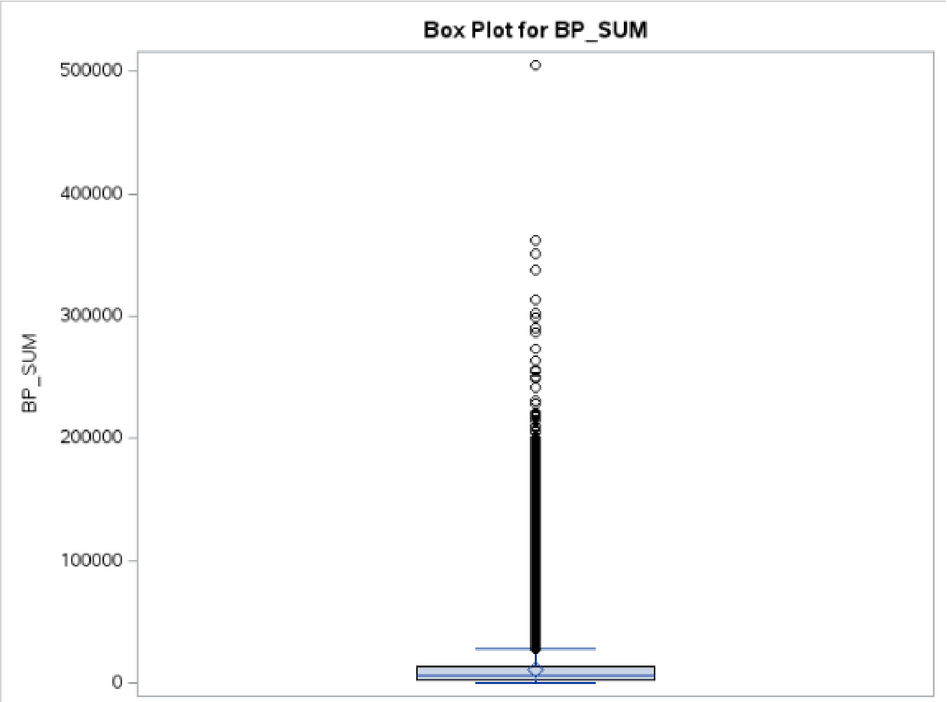


Figure 3.12.1: Box Plot of BP_SUM

Figure 3.12.1 shows that the box plot of BP_SUM includes some significant number of outliers represented by points above the whiskers. These outliers' values are higher than the majority of data points, it means that the values do not conform to the general pattern of BP_SUM.

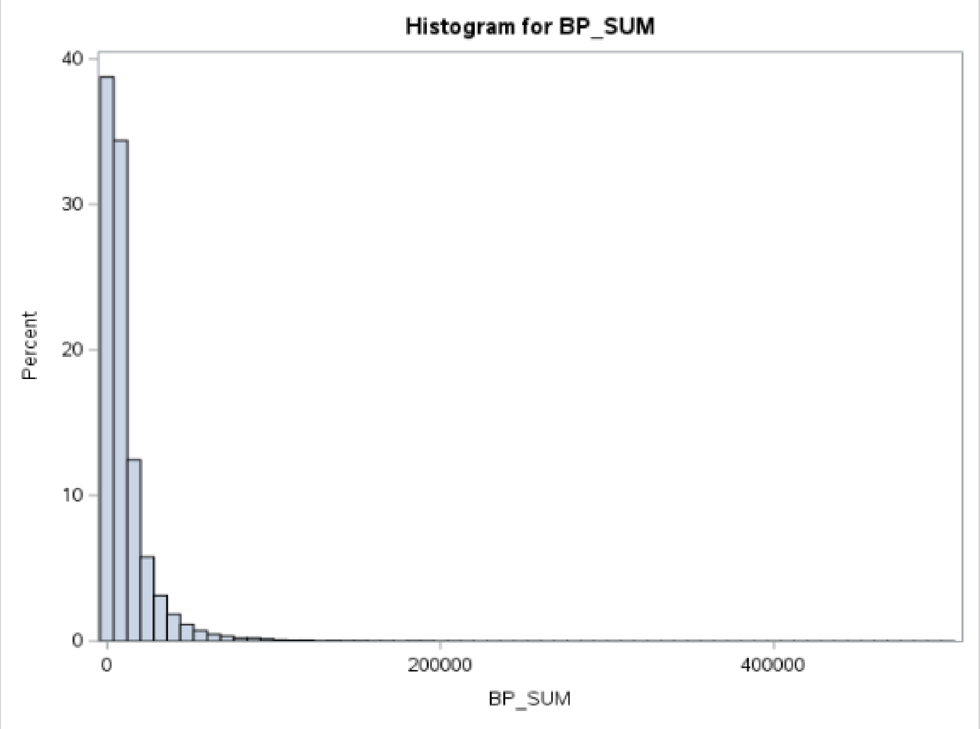


Figure 3.12.2: Histogram of BP_SUM

Figure 3.12.2 shows the Histogram of BP_SUM before any pre-processing, most of the values are concentrated at the lower end of scale, with a long tail extending to higher values. This presence of this long tail is indicative outliers in the dataset.

Pre-Processing Method:

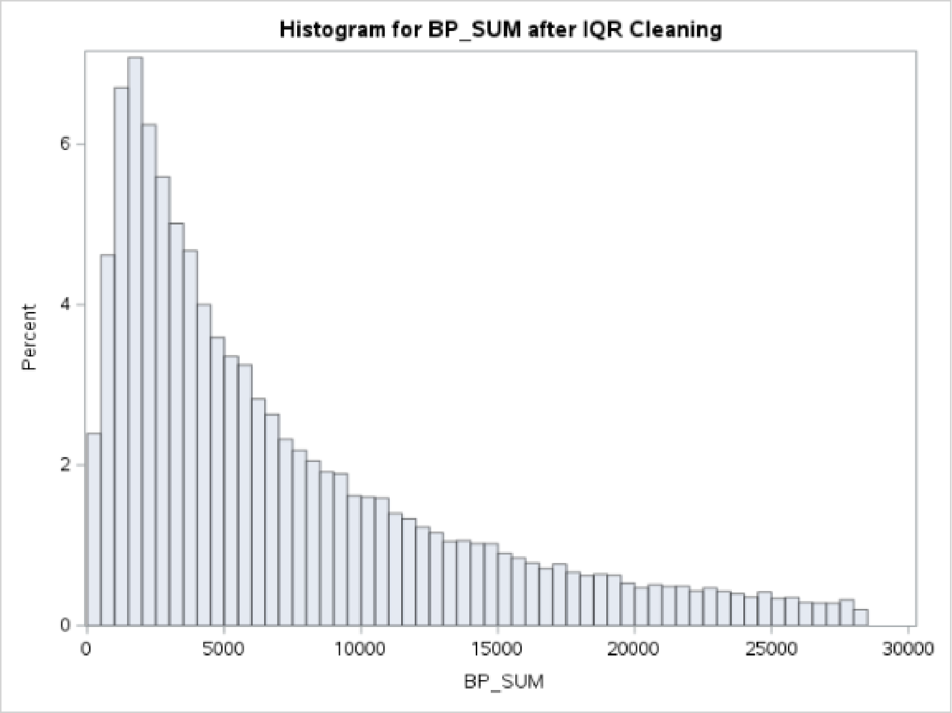


Figure 3.12.3: Histogram of BP_SUM after IQR Cleaning

Cleaning Method: Interquartile Range (IQR) Method for removing Outliers.

Figure 3.12.3 is about the after using IQR Cleaning in BP_SUM. By using this method, it will calculate the range of BP_SUM between the first quartile (25 percentage) and the third quartile (75 percentage). In the BP_SUM, if the data is above third quartile more than 1.5 times or below quartile 1.5 times will be considered as outlier and will be removed.

After Pre-Processing:

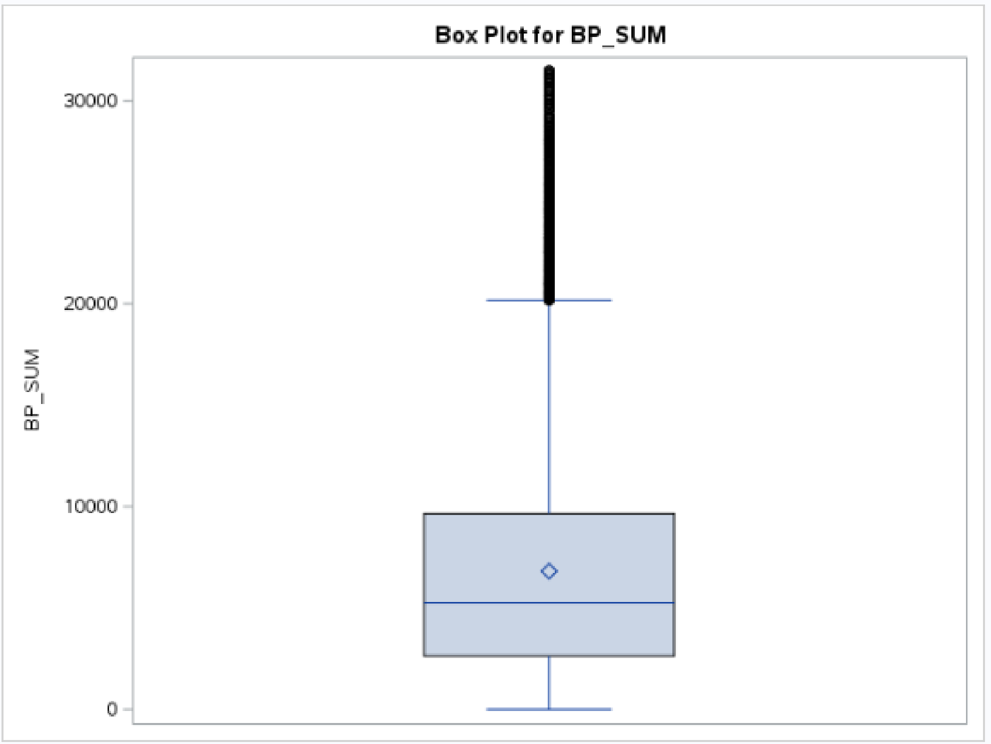


Figure 3.12.4: Box Plot of BP_SUM

Figure 3.12.4 represents the updated box plot which after using IQR method to clean the outliers. After cleaning the outliers, it shows that the data is now more tightly grouped, and the majority of values falling within narrower range.

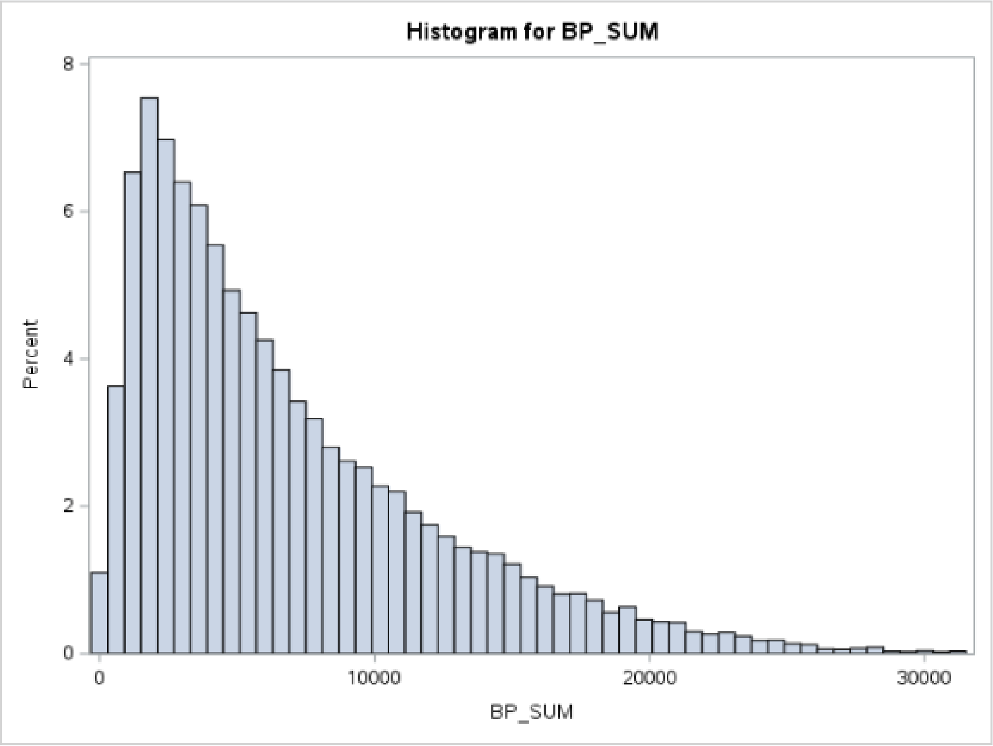


Figure 3.12.5: Histogram of BP_SUM

For this histogram of BP_SUM that shows in Figure 3.12.5, it shows that the extremely high values have been removed, and it is resulting the dataset that better represents the typical BP_SUM pattern. Now, the majority of values are closer to each of the data, this indicates a cleaner dataset by removing the outliers.

3.13 SUM_YR_1

Before Pre-Processing:

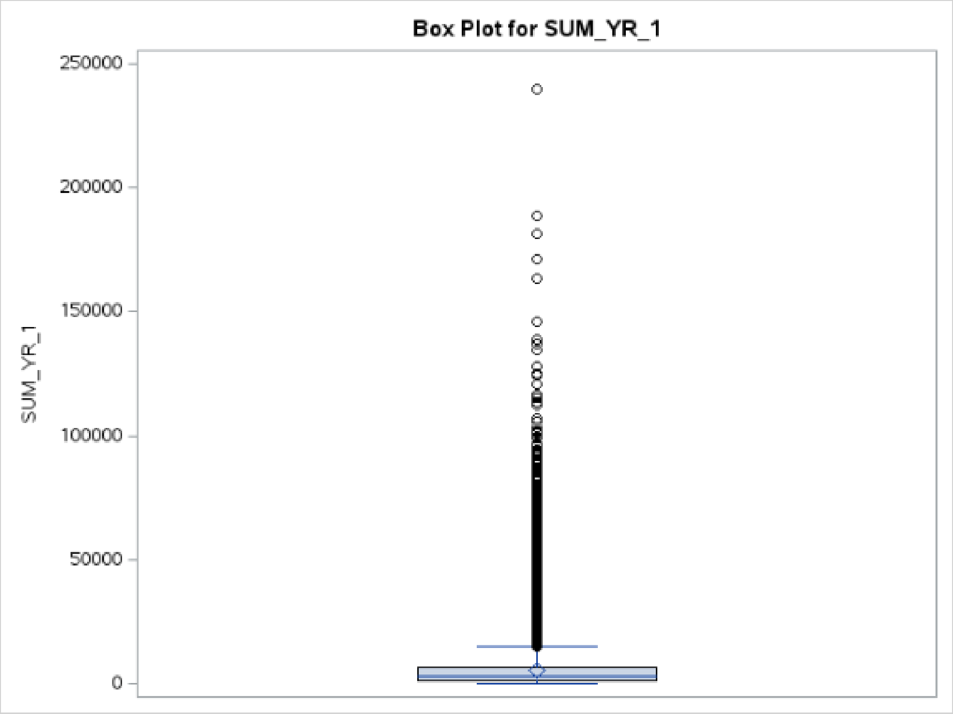


Figure 3.13.1: Box Plot of SUM_YR_1

The box plot of SUM_YR_1 has showed the large number of outliers represented above the range line. These outliers are significantly higher than the majority of the data points, indicating that some values do not follow the general pattern of SUM_YR_1.

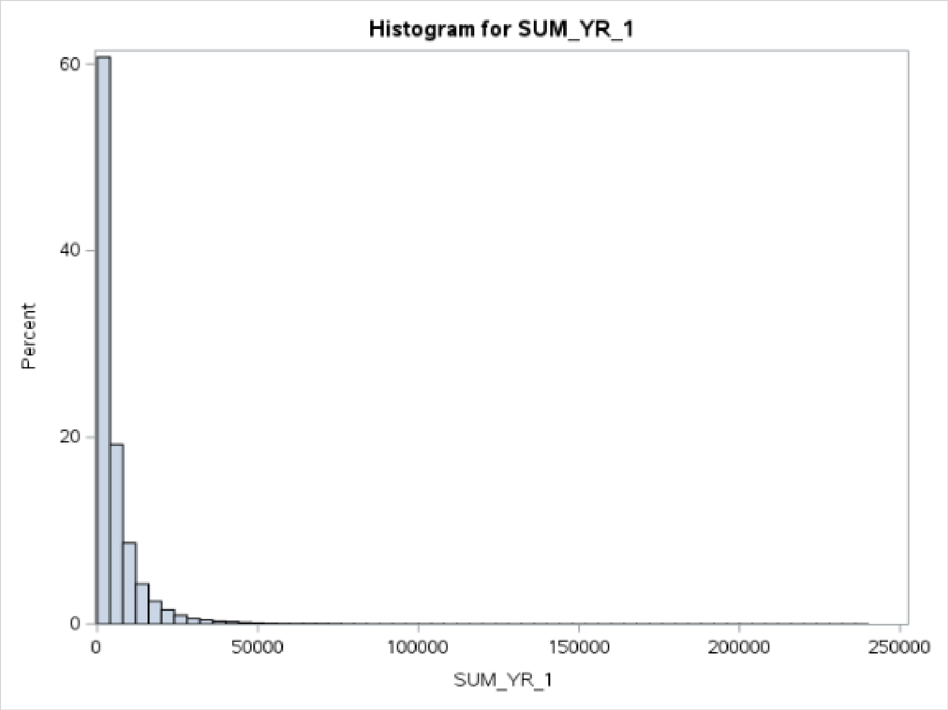
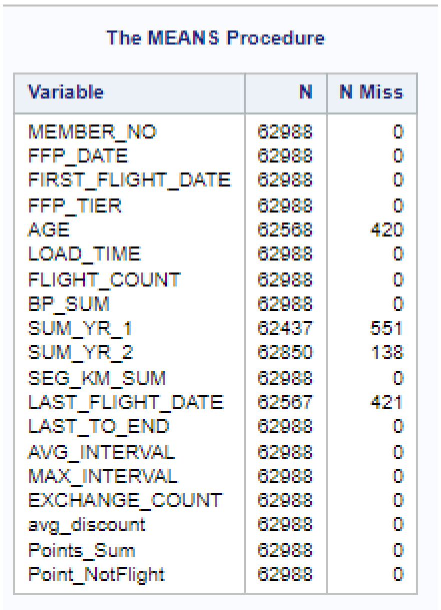


Figure 3.13.2: Histogram of SUM_YR_1

In this histogram of SUM_YR_1, it illustrates the distribution before implementing any pre-processing methods. The majority of values are clustered towards the lower end, while a significant number of high values create a long tail. This extended tail indicates the presence of outliers in the dataset.

Pre-Processing Method:



The MEANS Procedure

Variable	N	N Miss
MEMBER_NO	62988	0
FFP_DATE	62988	0
FIRST_FLIGHT_DATE	62988	0
FFP_TIER	62988	0
AGE	62568	420
LOAD_TIME	62988	0
FLIGHT_COUNT	62988	0
BP_SUM	62988	0
SUM_YR_1	62437	551
SUM_YR_2	62850	138
SEG_KM_SUM	62988	0
LAST_FLIGHT_DATE	62567	421
LAST_TO_END	62988	0
AVG_INTERVAL	62988	0
MAX_INTERVAL	62988	0
EXCHANGE_COUNT	62988	0
avg_discount	62988	0
Points_Sum	62988	0
Point_NotFlight	62988	0

Figure 3.13.3: Check for the missing values of SUM_YR_1

Cleaning Method: Checked for missing values.

Firstly, this pre-processing method had included for checking the missing values of SUM_YR_1. By checking the missing values, we used The MEANS Procedure to check the total missing values of every attribute and can clearly see that there are a total of 551 missing values in SUM_YR_1.

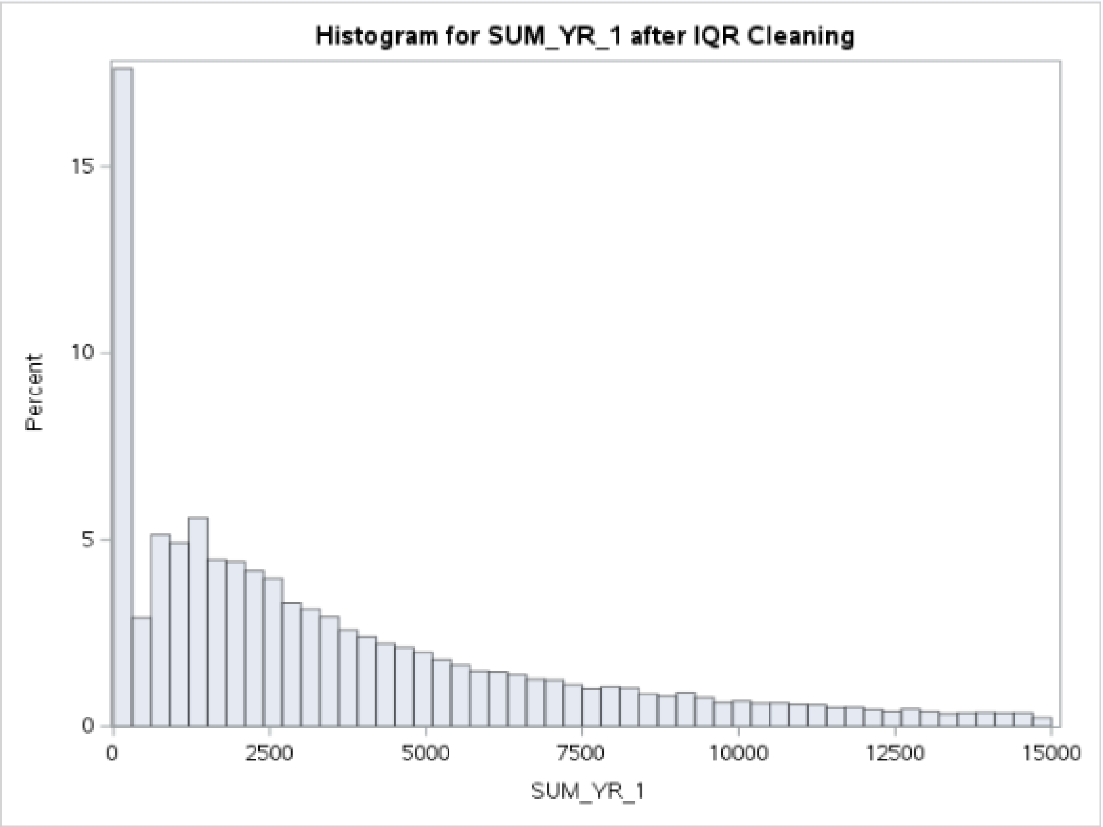


Figure 3.13.4: Histogram of SUM_YR_1 after IQR Cleaning

Cleaning Method: Interquartile Range Method for removing Outliers

Secondly, to remove the outliers which are performed not normally in the dataset pattern, we used the Interquartile Range (IQR) method to remove the outliers. By using this method, will involve calculating the IQR, which the range is between the first quartile and the third quartile of the data. When cleaning, it will check all the data which if it is 1.5 times the third quartile or 1.5 times the first quartile, it will be considered as outlier and will be removed in the dataset.

After Pre-Processing:

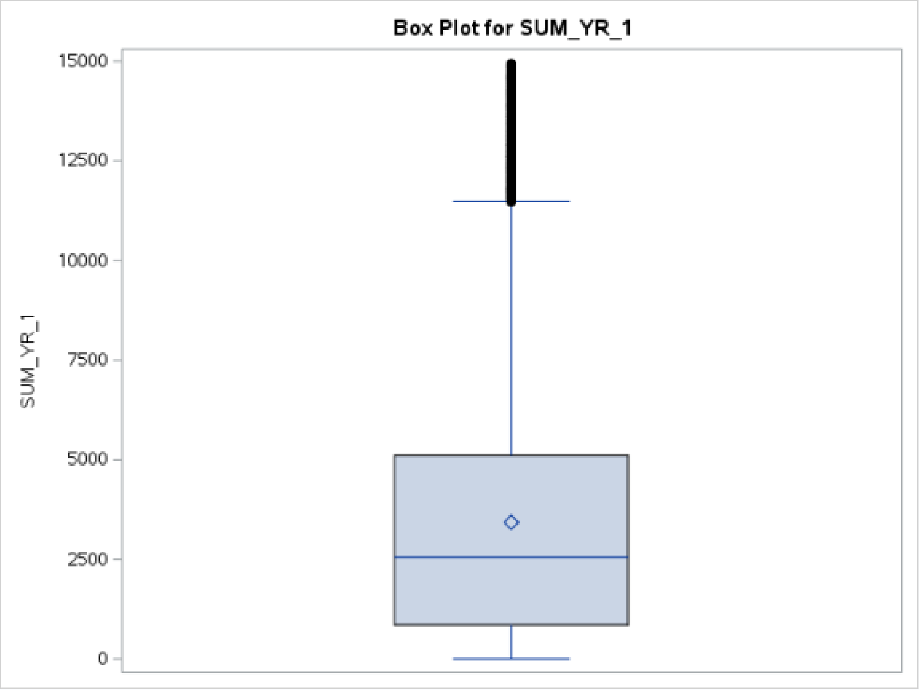


Figure 3.13.5: Box Plot of SUM_YR_1

In Figure 3.13.5, can clearly see that the box plot of after implement IQR cleaning method the outliers is fewer compare than the previous box plot in Figure 3.13.1.

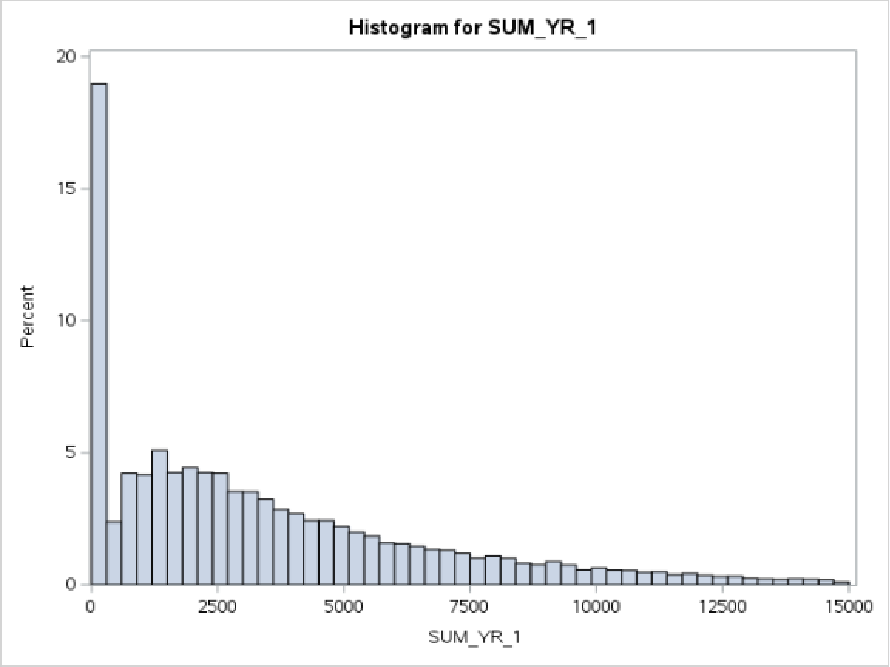


Figure 3.13.6: Histogram of SUM_YR_1

For Histogram, after implementing the IQR cleaning, in Figure 3.13.6 shows that the outliers has been removed and handling of missing value have resulted in a dataset more suitable for further analysis and modelling.

The application of the IQR method resulted in a dataset with a more consistent pattern, crucial for accurate analysis and reliable results in subsequent stages of data processing and modelling. The updated box plot and histogram visually confirm the effectiveness of these pre-processing steps.

3.14 SUM_YR_2

Before-Processing:

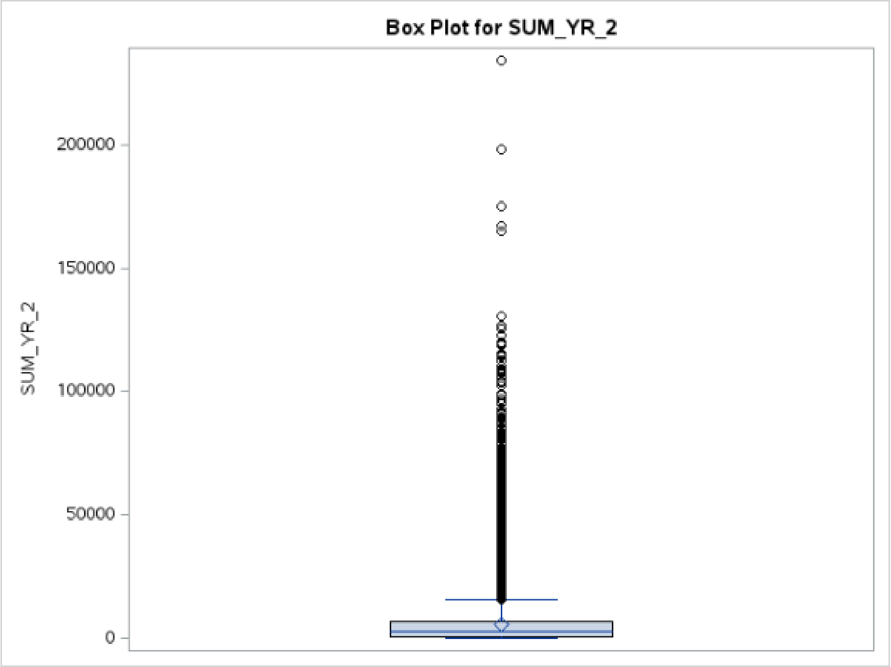


Figure 3.14.1: Box Plot of SUM_YR_2

The box plot of SUM_YR_2 shows a significant number of outliers, indicated by dotes far above the main distribution of data.

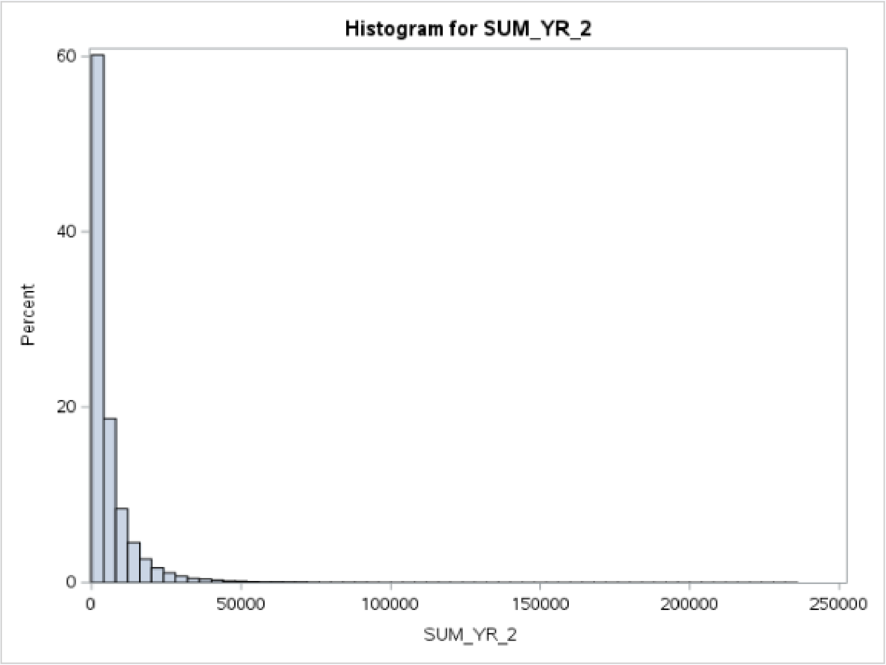


Figure 3.14.2: Box Plot of SUM_YR_2

The histogram of SUM_YR_2 reveals that the majority of the values are concentrated at the lower end of the histogram, with a long tail extending towards higher values. This indicates the presence of outliers in the dataset.

Pre-Processing Method:

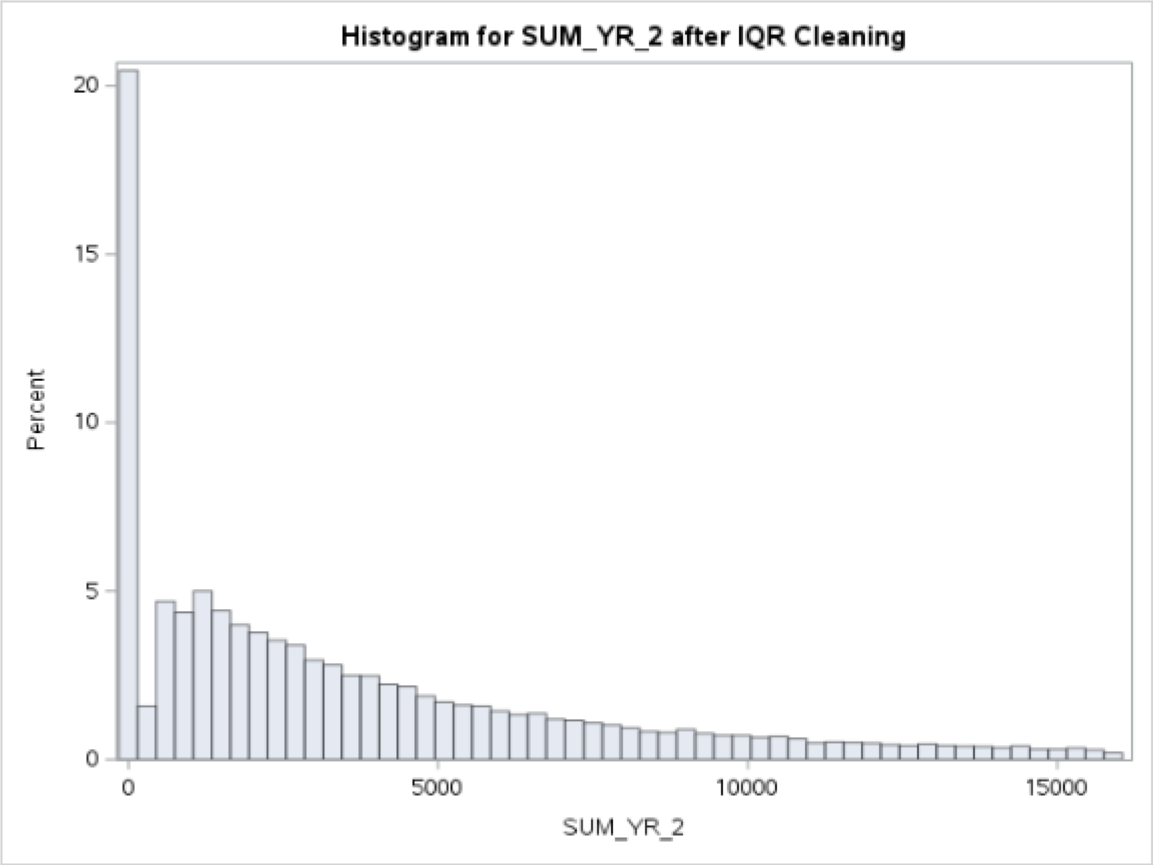


Figure 3.14.3: Histogram of SUM_YR_2 after IQR Cleaning

Cleaning Method: Interquartile Range Method for removing Outliers
To address the outliers in the SUM_YR_2 attribute, the IQR method will be used. The IQR method involves calculation in the first (Q1) and third (Q3) quartiles, and identifying outliers as any data points that fall below Q1 – 1.5/QR or above Q3 + 1.5/QR. This method is effectively in identifying and removing extreme values that could skew in the analysis.

After Pre-Processing:

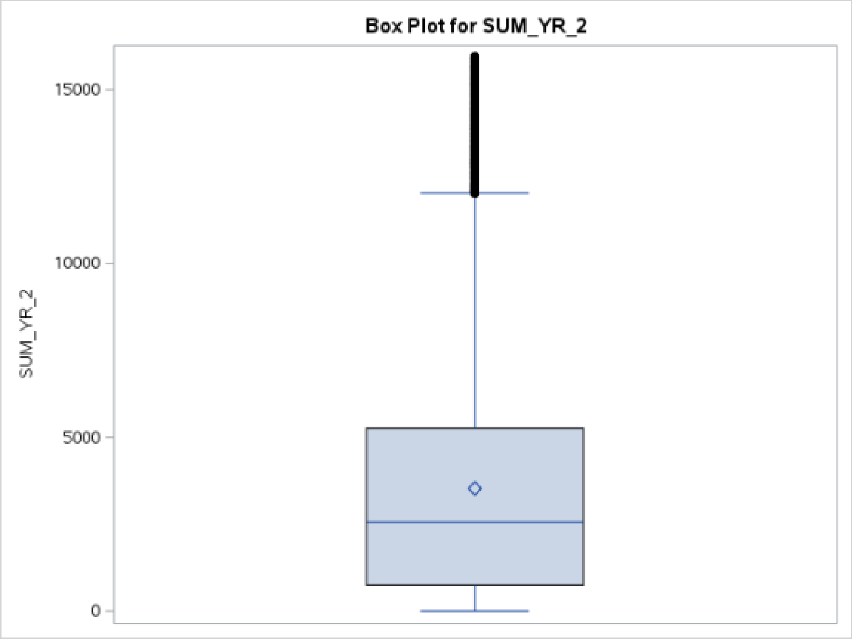


Figure 3.14.4: Box Plot of SUM_YR_2

The box plot of SUM_YR_2 post-cleaning also reflects this change, with fewer outliers present.

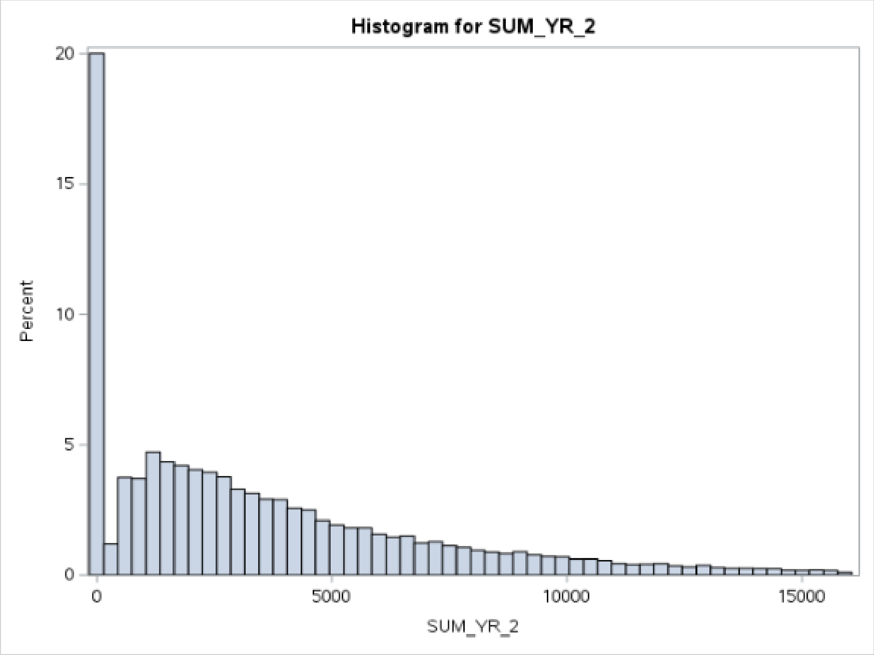


Figure 3.14.5: Histogram of SUM_YR_2

The histogram of SUM_YR_2 after IQR cleaning shows a more compact distribution of data, with the extreme values removed.

3.15 SEG_KM_SUM

Before Pre-Processing:

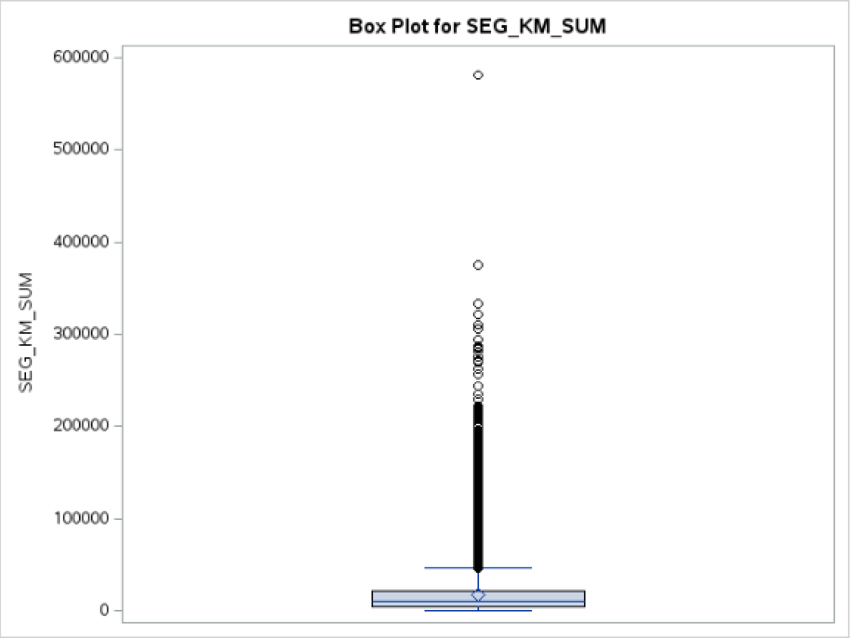


Figure 3.15.1: The Box Plot of SEG_KM_SUM

In this boxplot of SEG_KM_SUM, it displays there are many of extreme outliers existed, and is indicating a wide range of values. By the appearance of outliers had skewed the box plot and makes the actual data distribution difficult to interpret.

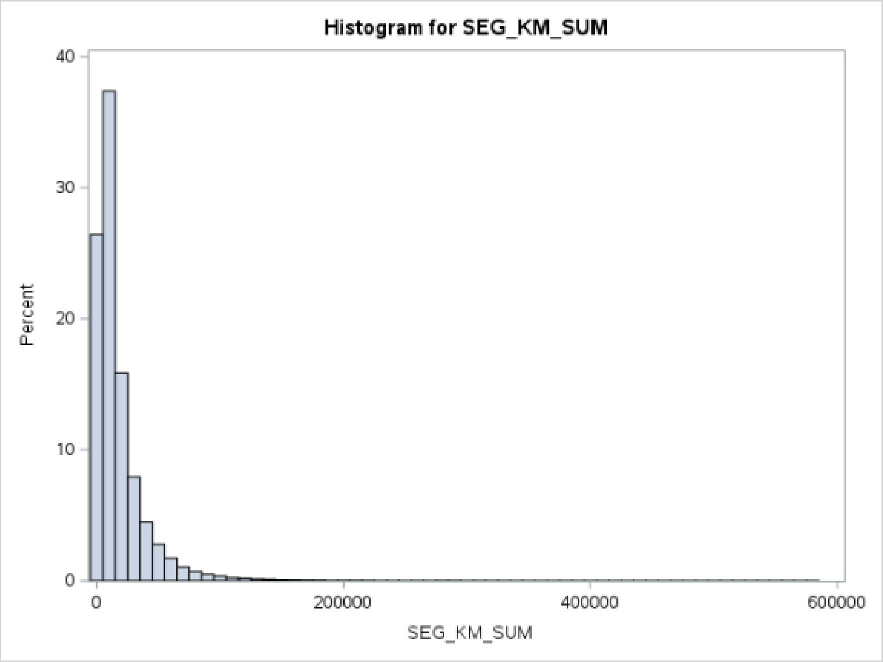


Figure 3.15.2: The Histogram of SEG_KM_SUM

Besides than box plot, in the histogram can also clearly see that outliers existed. Is shows a highly right-skewed distribution with a long tail. Besides, a large concentration of data points at the lower end, with few extremely high values can be observe in the Figure 3.15.2.

Pre-Processing Method:

Cleaning Method: IQR Method for Outliers

My using Interquartile (IQR) Method, needs to be calculate the first quartile(Q1) and third quartile (Q3). Next, need to compute the IQR, which the formula is IQR = Q3 - Q1. Thirdly, by using this method, needs to define the lower and upper bounds these are the formula:

"lower bound = Q1 – 1.5*IQR" and "upper bound = Q3 + 1.5*IQR"

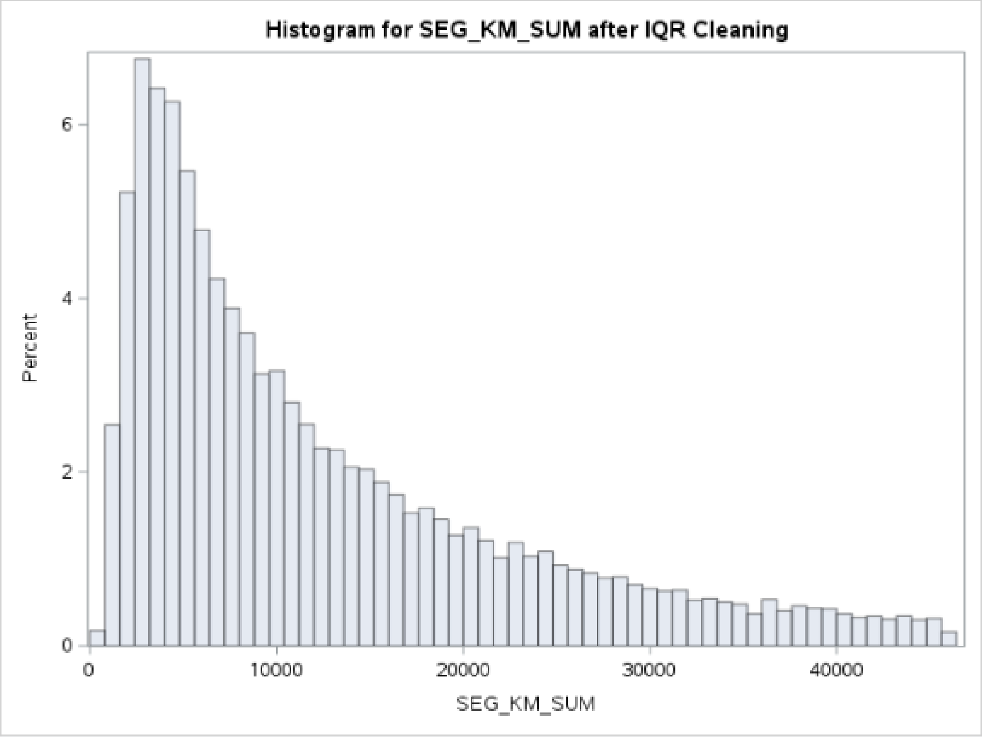


Figure 3.15.3: Histogram of SEG_KM_SUM after IQR Cleaning

After cleaning, the Histogram of SEG_KM_SUM now shows a more normal-like distribution with reduced skewness. The extreme values beyond the upper bound are removed, leading to a clearer view of the central data distribution.

After Pre-Processing:

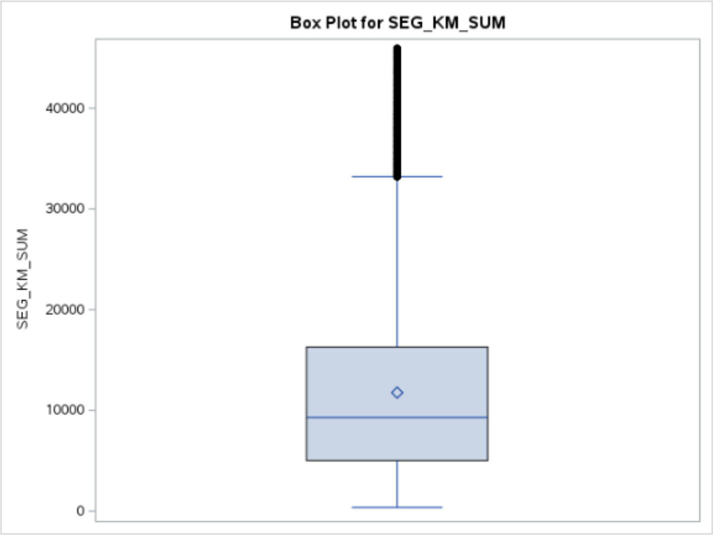


Figure 3.15.4: Updated Box Plot of SEG_KM_SUM

The updated box plot that can be seen in Figure 3.15.4 becomes more representative of the central data distribution. Outliers in this previous boxplot are significantly reduced, providing a clearer view of the median, quartiles and overall spread of the data.

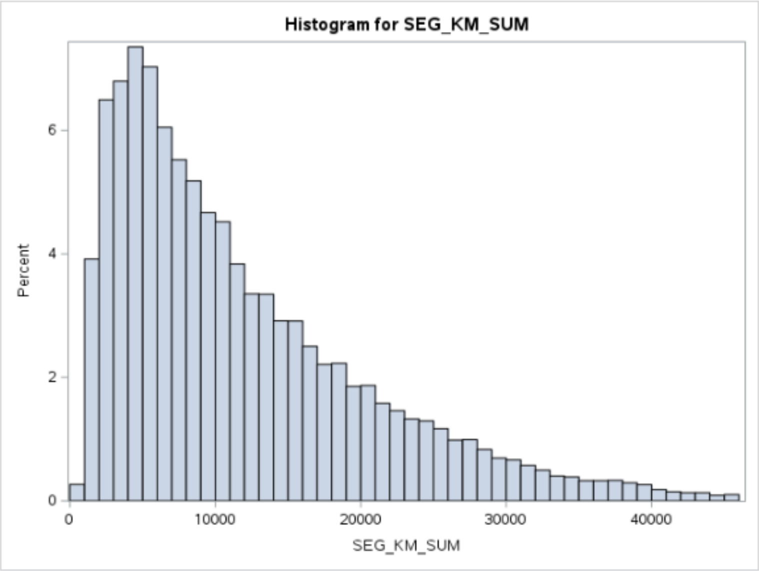


Figure 3.15.5: Updated Histogram of SEG_KM_SUM

The updated histogram of SEG_KM_SUM shows a more spread distribution without extreme values. Now the updated histogram's tail is shorter, and the central tendency of the data is more apparent.

3.16 LAST_FLIGHT_DATE

Before Pre-Processing:

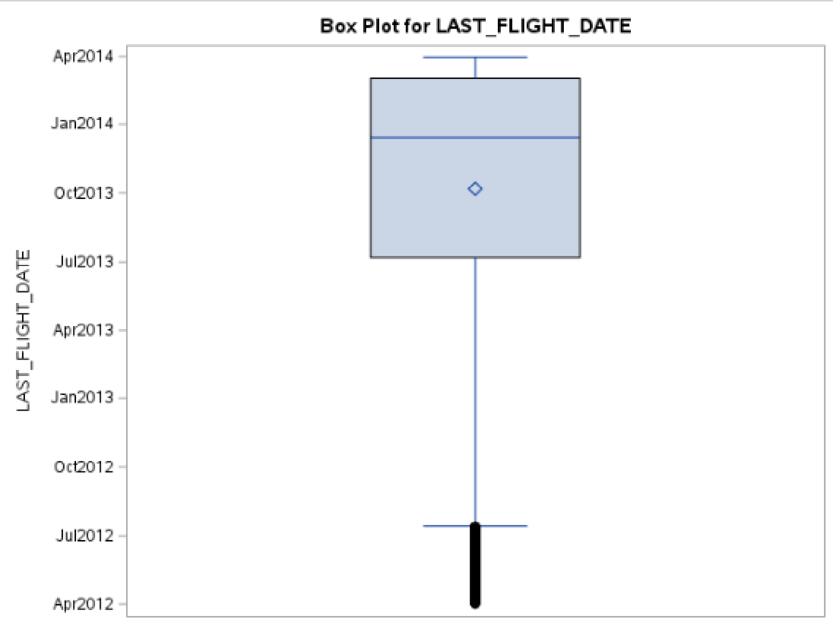


Figure 3.16.1: Box Plot of LAST_FLIGHT_DATE

The box plot highlights several outliers in the dataset, which lower than the normal range. These outliers indicate the flight dates that are significantly earlier than the majority of the data.

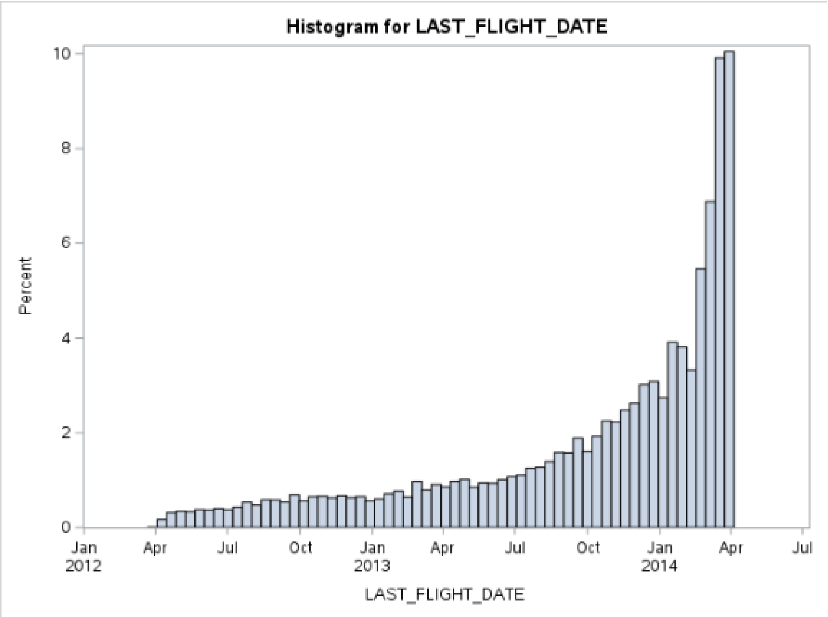
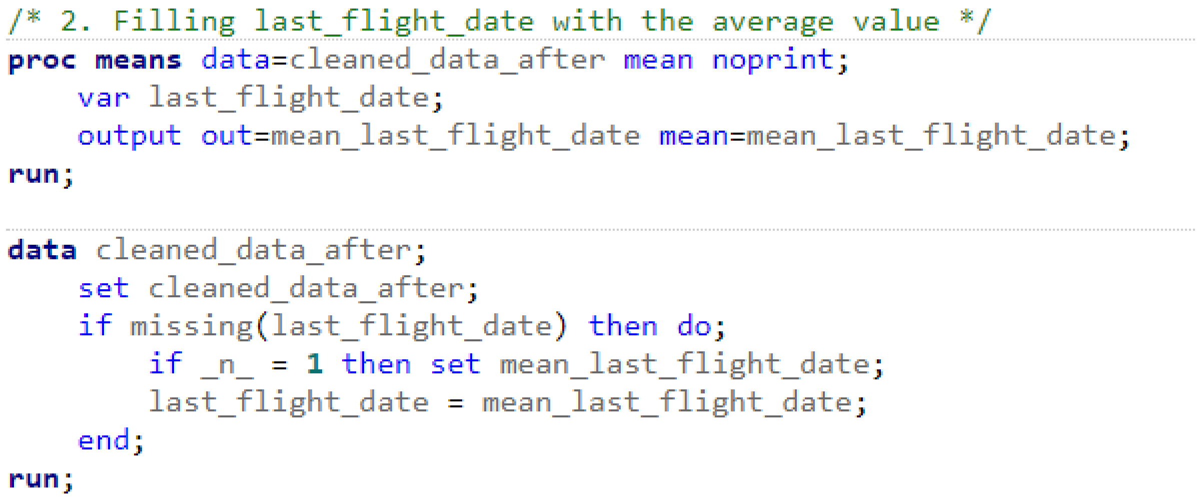


Figure 3.16.2: Histogram of LAST_FLIGHT_DATE

The histogram displays an increasing frequency trend over time, with a notable rise in recent months.

Pre-Processing Method:



```
/* 2. Filling last_flight_date with the average value */
proc means data=cleaned_data_after mean noprint;
    var last_flight_date;
    output out=mean_last_flight_date mean=mean_last_flight_date;
run;

data cleaned_data_after;
    set cleaned_data_after;
    if missing(last_flight_date) then do;
        if _n_ = 1 then set mean_last_flight_date;
        last_flight_date = mean_last_flight_date;
    end;
run;
```

Figure 3.16.3: Code to replace missing values with the average value

Cleaning Method: Replace missing value using average value.

In the LAST_FLIGHT_DATE attribute, there are a total of 421 data missing. The missing values in the LAST_FLIGHT_DATE attribute will be filled using average value. This is to ensure the dataset remains complete and avoids potential biases caused by missing data.

In Figure 3.16.3, this code snippet demonstrates how the average value was calculated and used to replace missing entries in the LAST_FLIGHT DATE attribute.

After Pre-Processing:

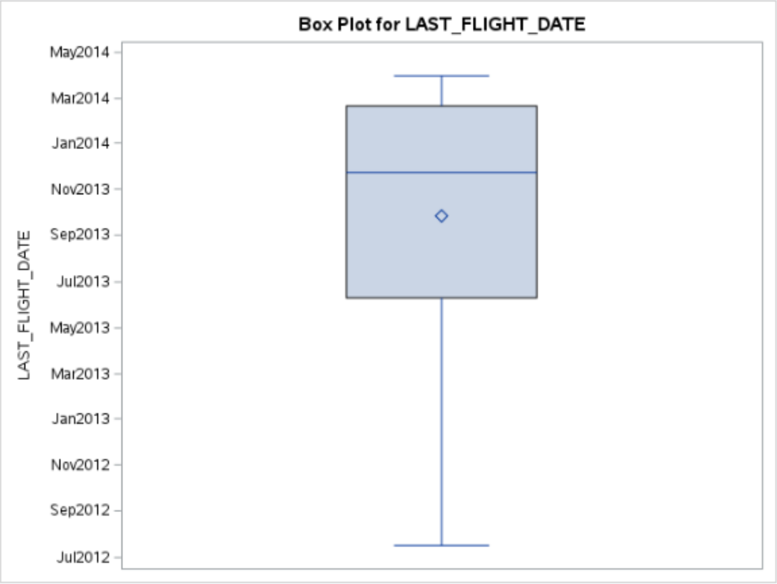


Figure 3.16.4: The updated Box Plot of LAST_FLIGHT_DATE

The revised box plot shows a more consistent data distribution after filling missing values. Can clearly see that the outliers are less shown, and the central tendency is clearer.

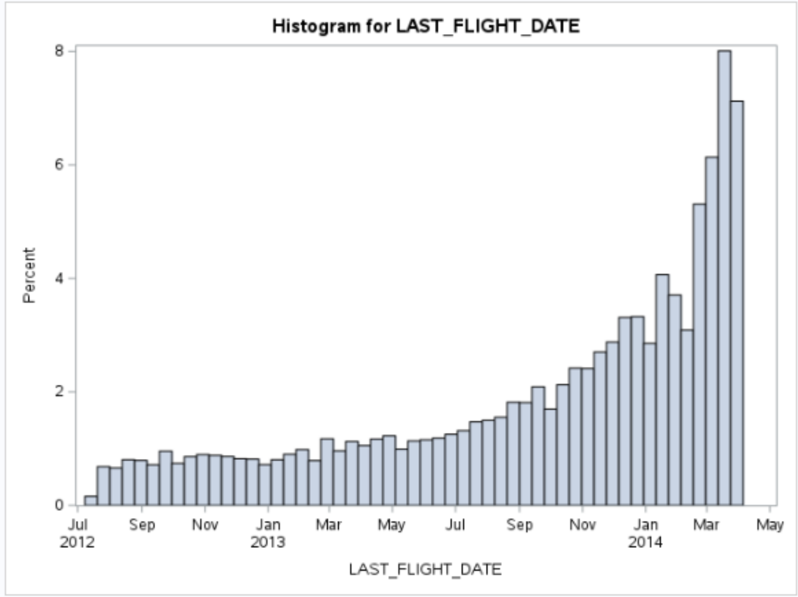


Figure 3.16.5: The updated Histogram of LAST_FLIGHT_DATE

The updated histogram reflects the complete dataset, providing a clearer picture of flight date distribution. The trend of increasing frequency over time is more pronounced, highlighting the dataset’s temporal characteristics.

3.17 LAST_TO_END

Before Pre-Processing:

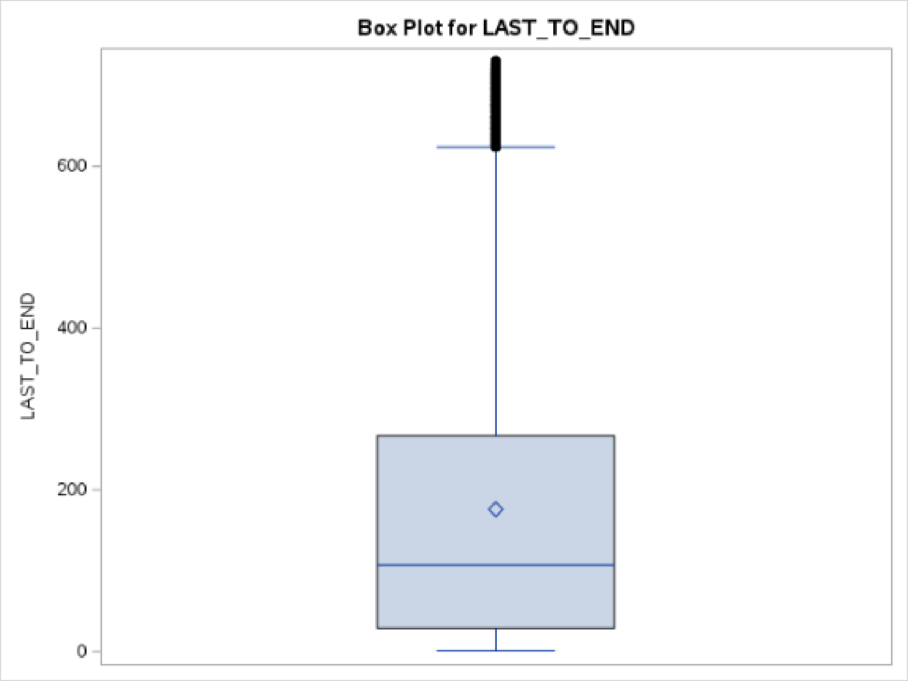


Figure 3.17.1: Box Plot of LAST_TO_END

There are a lot of outliers at the upper end of the range, as box plot shows. With a lengthy tail that extends beyond 300, the bulk of the data points are clustered below this value. There may be a large number of extreme values or outliers in the dataset if there are multiple data points outside the top whisker.

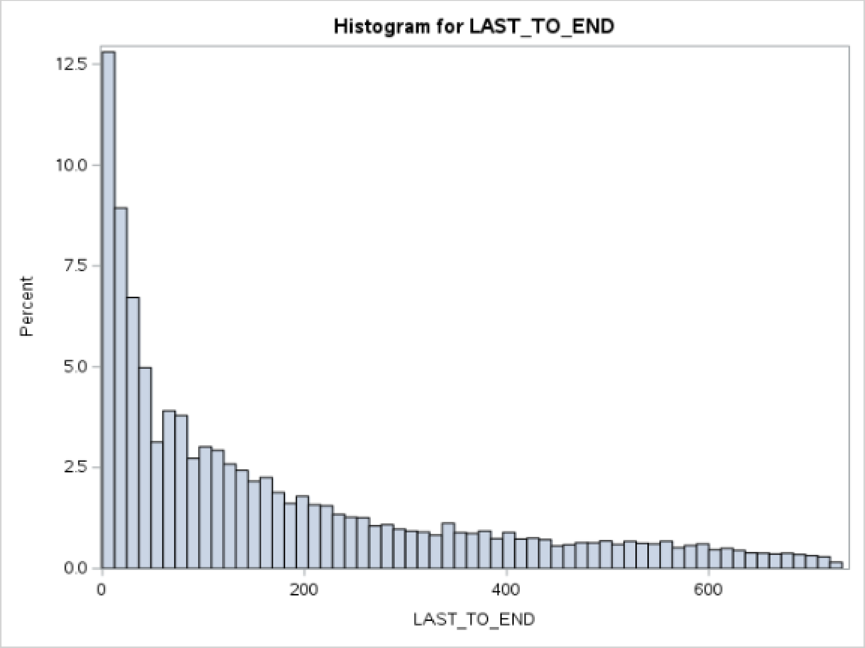
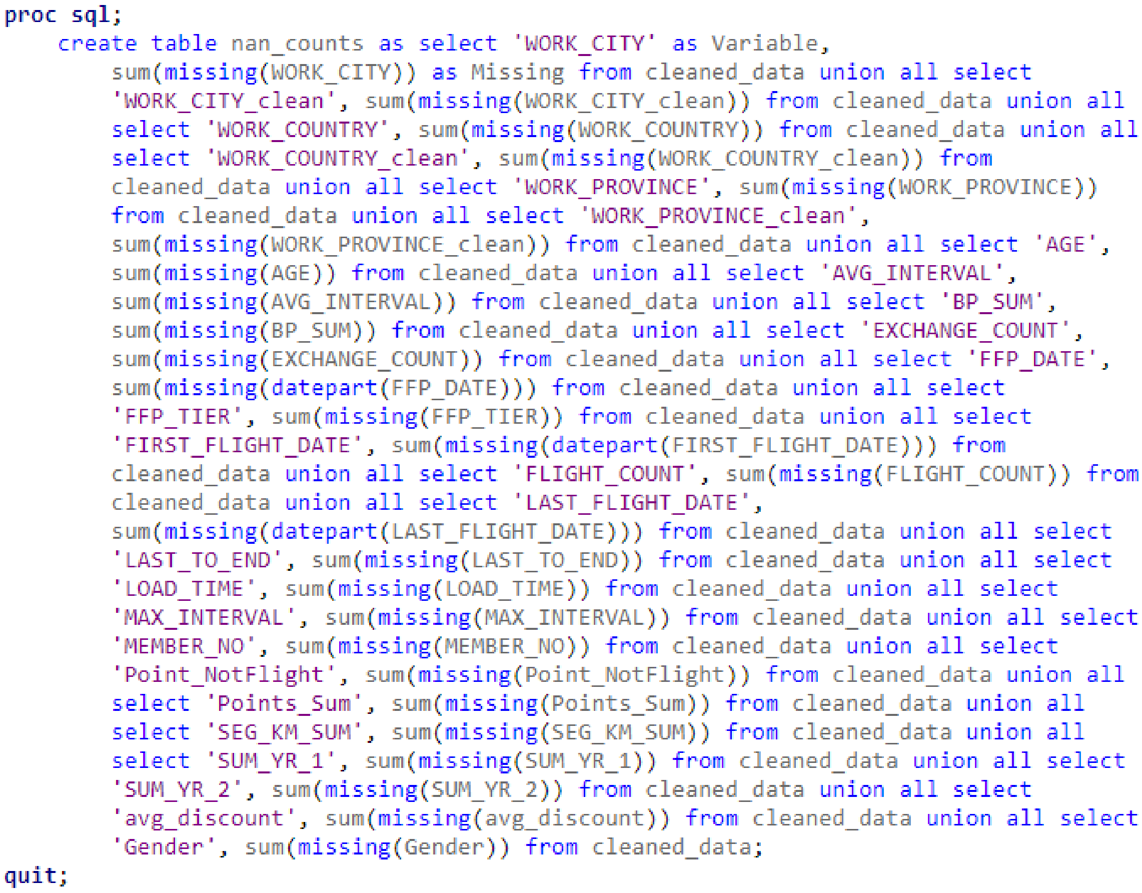


Figure 3.17.2: Histogram of LAST_TO_END

In Figure 3.17.2, the histogram displays a positively skewed distribution of LAST_TO_END values. In the figure 3.17.2, can clearly see that most of the values clustered at the lower end of the range, with the highest frequency occurring near zero. The frequency diminished as the value increased, forming a long tail that extends towards higher values. This long tail further corroborates the presence of outliers in the dataset.

Pre-Processing Method:

Cleaning Method: Checked for missing values.



```
proc sql;
    create table nan_counts as select 'WORK_CITY' as Variable,
        sum(missing(WORK_CITY)) as Missing from cleaned_data union all select
        'WORK_CITY_clean', sum(missing(WORK_CITY_clean)) from cleaned_data union all
        select 'WORK_COUNTRY', sum(missing(WORK_COUNTRY)) from cleaned_data union all
        select 'WORK_COUNTRY_clean', sum(missing(WORK_COUNTRY_clean)) from
        cleaned_data union all select 'WORK_PROVINCE', sum(missing(WORK_PROVINCE))
        from cleaned_data union all select 'WORK_PROVINCE_clean',
        sum(missing(WORK_PROVINCE_clean)) from cleaned_data union all select 'AGE',
        sum(missing(AGE)) from cleaned_data union all select 'AVG_INTERVAL',
        sum(missing(AVG_INTERVAL)) from cleaned_data union all select 'BP_SUM',
        sum(missing(BP_SUM)) from cleaned_data union all select 'EXCHANGE_COUNT',
        sum(missing(EXCHANGE_COUNT)) from cleaned_data union all select 'FFP_DATE',
        sum(missing(datepart(FFP_DATE))) from cleaned_data union all select
        'FFP_TIER', sum(missing(FFP_TIER)) from cleaned_data union all select
        'FIRST_FLIGHT_DATE', sum(missing(datepart(FIRST_FLIGHT_DATE))) from
        cleaned_data union all select 'FLIGHT_COUNT', sum(missing(FLIGHT_COUNT)) from
        cleaned_data union all select 'LAST_FLIGHT_DATE',
        sum(missing(datepart(LAST_FLIGHT_DATE))) from cleaned_data union all select
        'LAST_TO_END', sum(missing(LAST_TO_END)) from cleaned_data union all select
        'LOAD_TIME', sum(missing(LOAD_TIME)) from cleaned_data union all select
        'MAX_INTERVAL', sum(missing(MAX_INTERVAL)) from cleaned_data union all select
        'MEMBER_NO', sum(missing(MEMBER_NO)) from cleaned_data union all select
        'Point_NotFlight', sum(missing(Point_NotFlight)) from cleaned_data union all
        select 'Points_Sum', sum(missing(Points_Sum)) from cleaned_data union all
        select 'SEG_KM_SUM', sum(missing(SEG_KM_SUM)) from cleaned_data union all
        select 'SUM_YR_1', sum(missing(SUM_YR_1)) from cleaned_data union all select
        'SUM_YR_2', sum(missing(SUM_YR_2)) from cleaned_data union all select
        'avg_discount', sum(missing(avg_discount)) from cleaned_data union all select
        'Gender', sum(missing(Gender)) from cleaned_data;
quit;
```

Figure 3.17.3: The code of checking missing values

Based on this code, it snippet checks for missing values in the LAST_TO_END attribute using the missing() function in a proc sql statement. This step is to ensure the dataset’s completeness and identifies any gaps that need addressing.

Cleaning Method: IQR Method for removing Outliers.

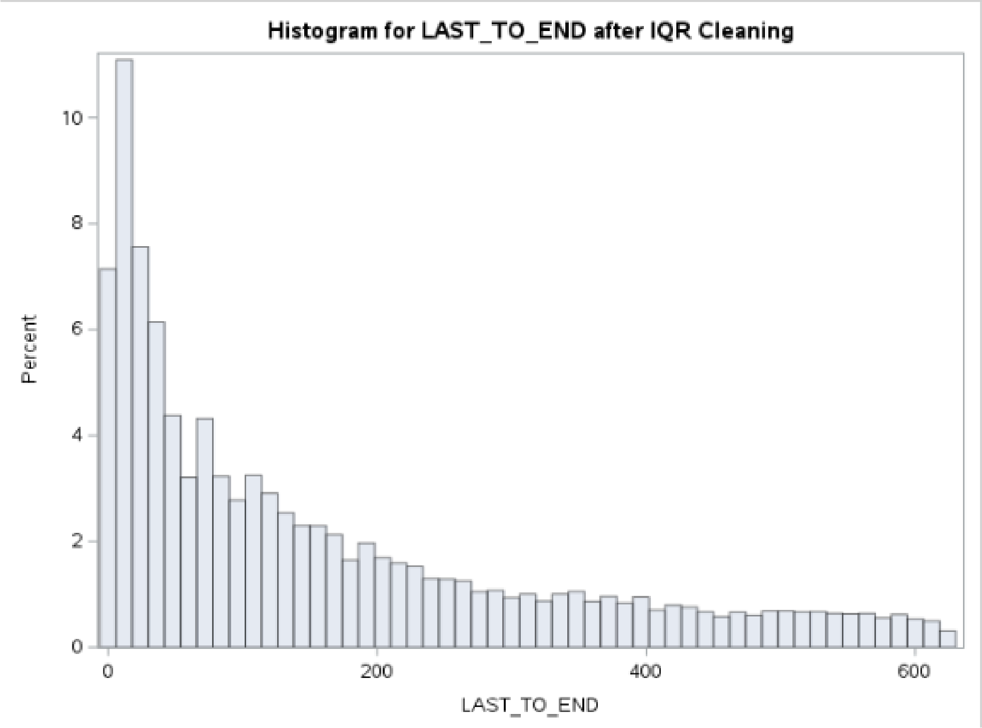


Figure 3.17.4: Histogram of LAST_TO_END after IQR Cleaning

The Interquartile Range (IQR) method was employed to identify and remove the outliers. This method involves calculating the IQR which was the ranges of first quartile (Q1) and the third quartile (Q3) and determine the boundaries beyond which data point is considered as outliers and remove the data from the column.

After Pre-Processing:

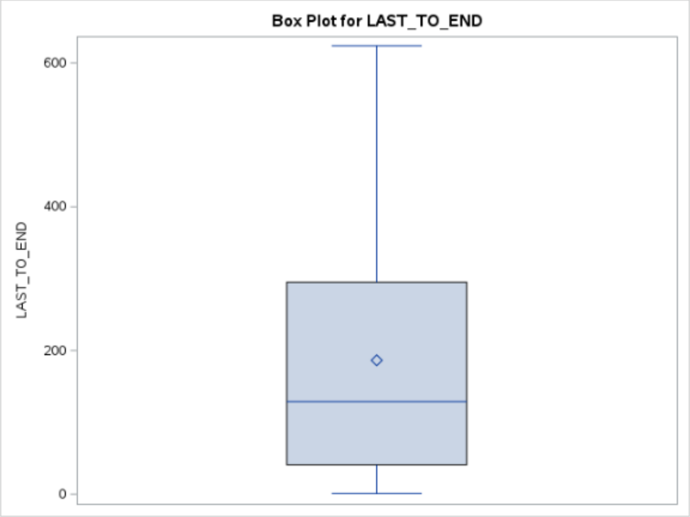


Figure 3.17.5: Updated Box Plot of LAST_TO_END

In updated box plot (Figure 3.17.5), it demonstrates a substantial reduction in outliers. The majority of the data points now fall within the interquartile range, and the whiskers extend to a more reasonable range, reflecting a cleaner dataset. Removing the extreme values (outliers) has resulted in a more accurate representation of the central tendency and variability of the LAST_TO_END attribute.

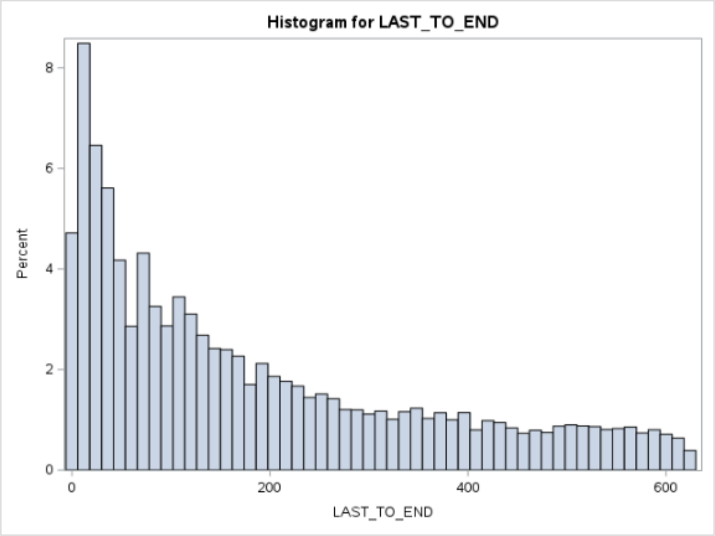


Figure 3.17.6: Updated Histogram of LAST_TO_END

The cleaned histogram which was Figure 3.17.6 has illustrates the improved data distribution. The peak near zero remains, but the distribution is more even and less skewed after cleaning. The frequency values decreased more steadily, and the presence of extreme values has been minimized, leading to a dataset that is more suitable for subsequent analysis.

3.18 AVG_INTERVAL

Before Pre-processing:

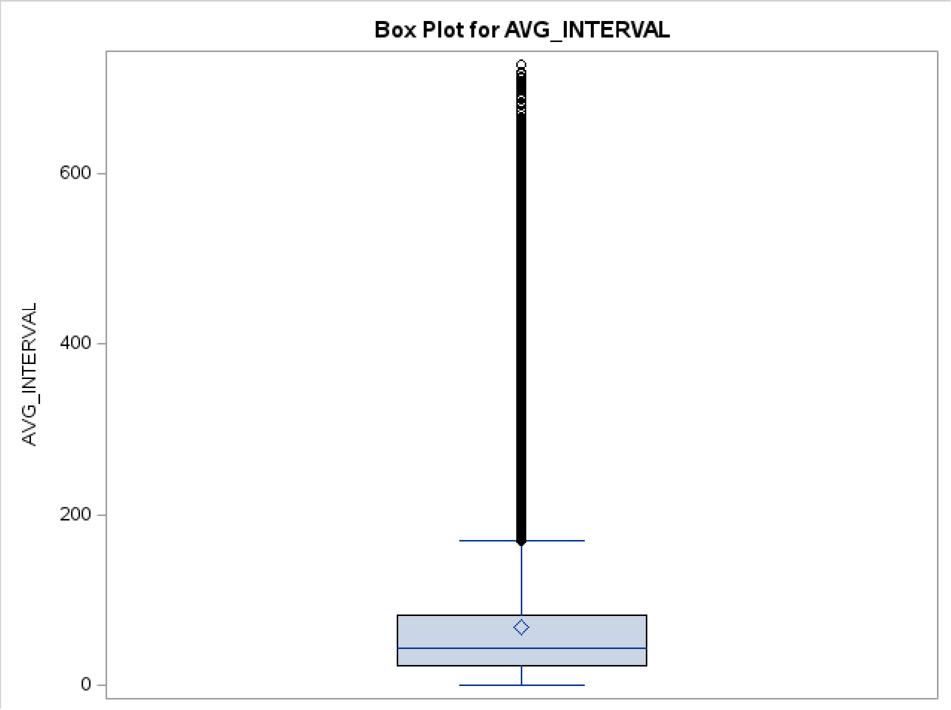


Figure 3.18.1: Box Plot of AVG_INTERVAL

In Figure 3.18.1, the box plot of AVG_INTERVAL shows a significant number of outliers extending far above the upper whisker. Most of the data points are concentrated near the lower end of the scale, indicating a right-skewed distribution.

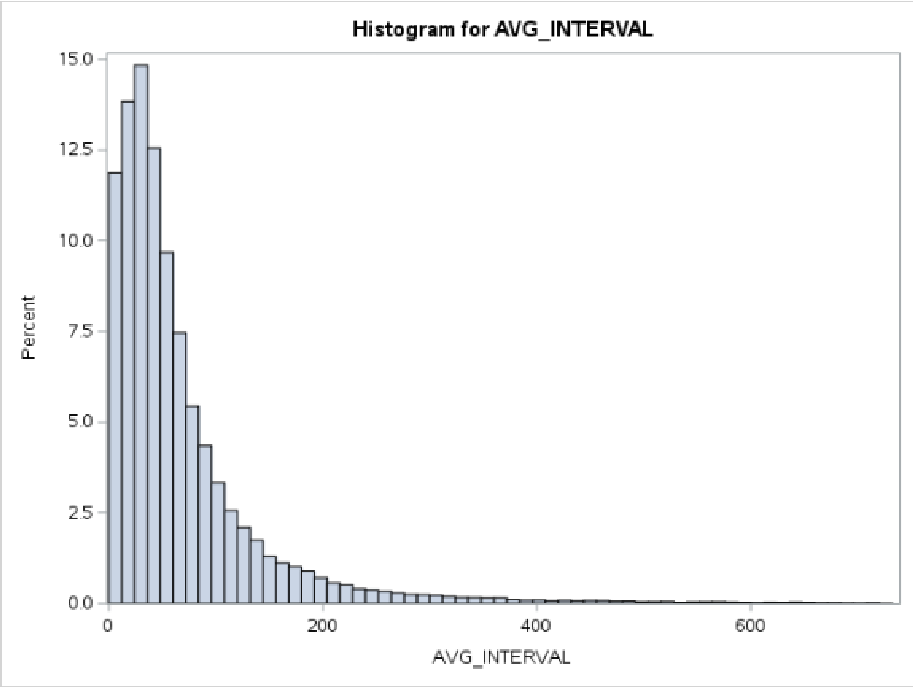


Figure 3.18.2: Histogram of AVG_INTERVAL

In the histogram of AVG_INTERVAL illustrates the right-skewed distribution values. A large portion of the data is clustered towards the lower end with a long tail extending to higher values. This distribution confirms the presence of numerous high-value outliers.

Pre-Processing Method:

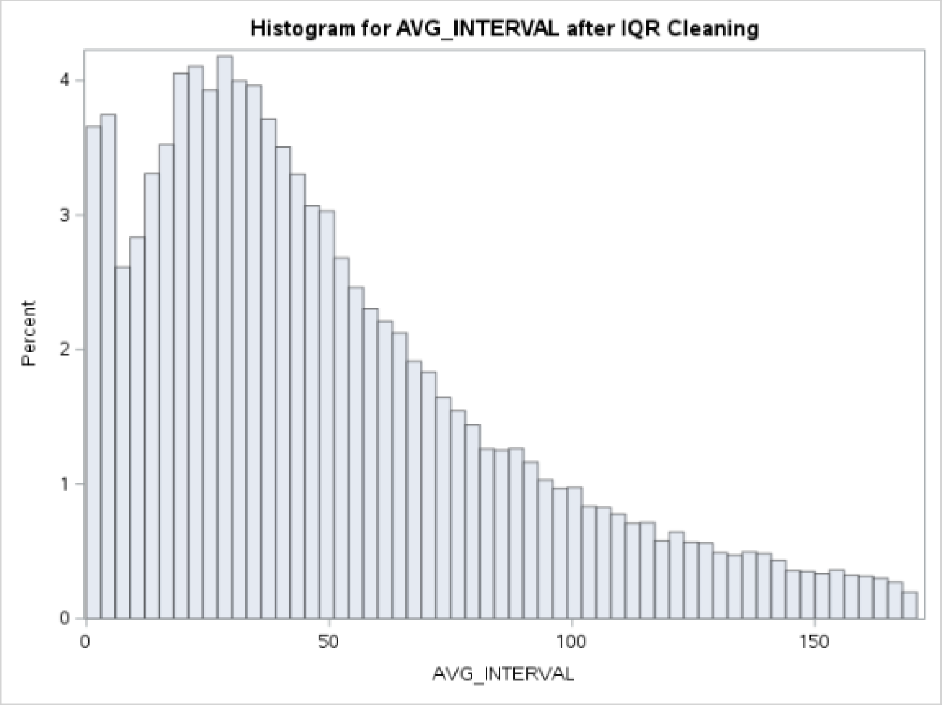


Figure 3.18.3: Histogram of AVG_INTERVAL after IQR Cleaning

Cleaning Method: IQR Method for removing Outliers.

In this cleaning method, IQR involves in calculating the IQR, afterwards will be identifying and excluding data points that lie beyond 1.5 times the IQR above the third quartile (Q3) or below the first quartile (Q1).

After Pre-Processing:

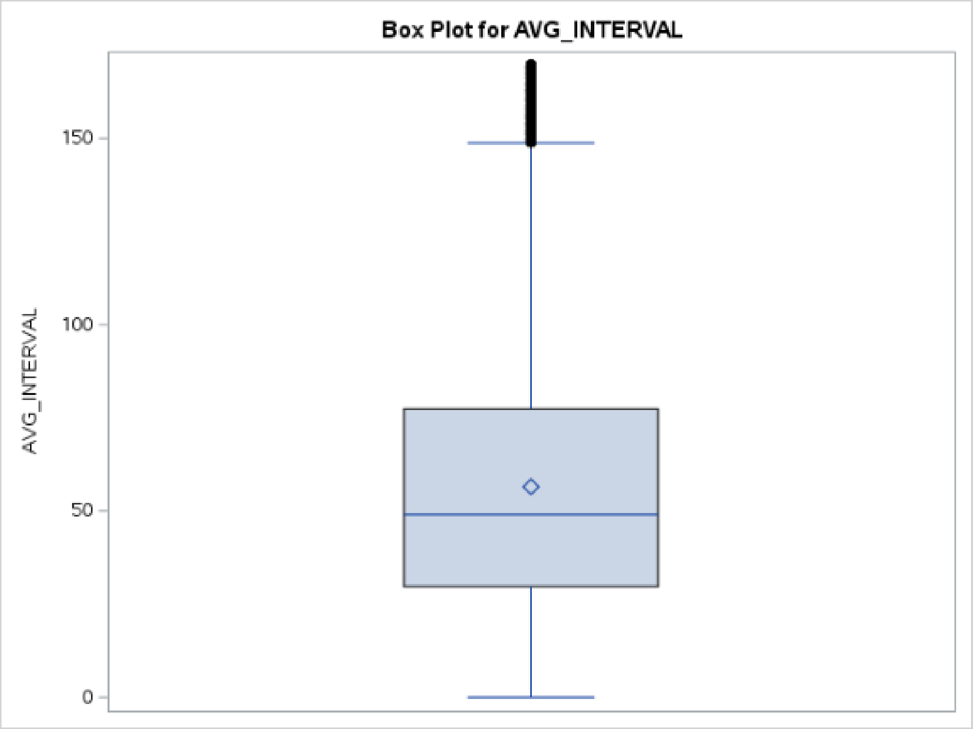


Figure 3.18.4: Updated Box Plot of AVG_INTERVAL

In the updated box plot, it reflects the fewer outliers above the whiskers, and the overall range of data is more compressed. Besides, the tendency and variability of AVG_INTERVAL data are now more accurately represented. The cleaning represents facilitates better understanding and analysis of the typical interview values in the dataset.

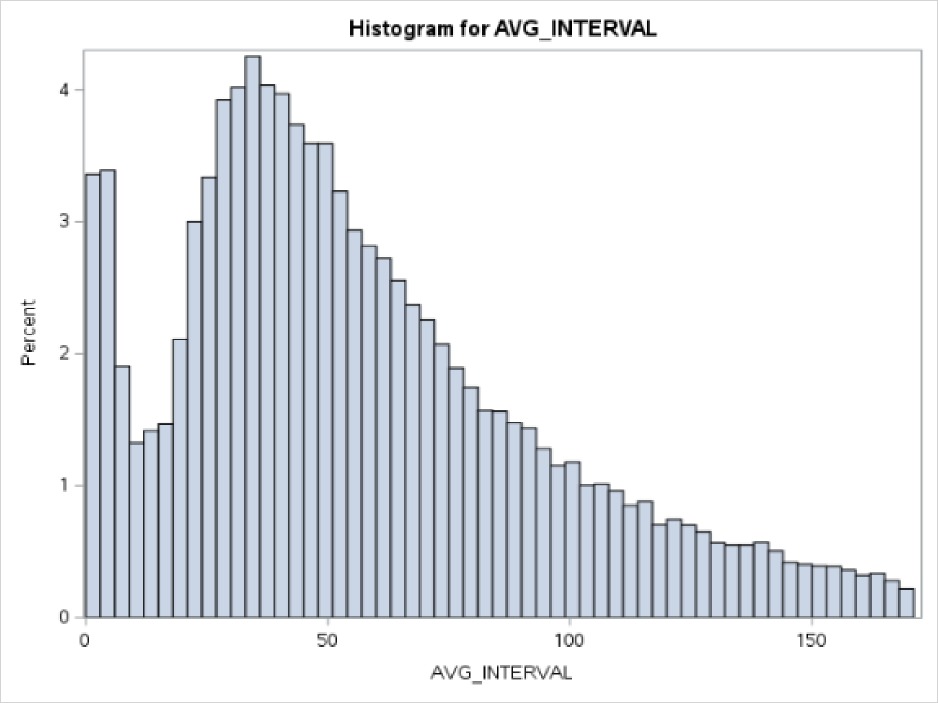


Figure 3.18.5: Updated Histogram of AVG_INTERVAL

In Figure 3.18.5 shows the updated histogram of AVG_INTERVAL and display a more balanced distribution value. The concentration of the data points around the median is more apparent with fewer extreme values distorting the overall distribution. This improvement enhances the dataset's suitability for further analysis and modelling.

3.19 MAX_INTERVAL

Before Pre-Processing:

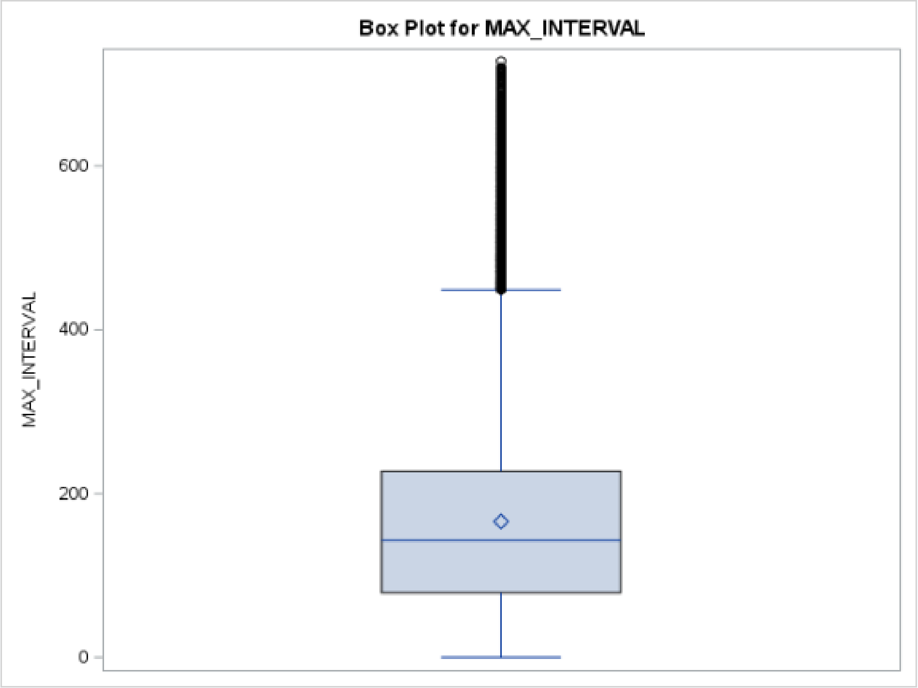


Figure 3.19.1: Box Plot of MAX_INTERVAL

In the Figure 3.19.1, there are substantial number of outliers at the upper end of the range. With a long tail that extends beyond 200, most data points are clustered below this value. There are also several data points outside the box plot's upper border, indicate that there may be a significant number of extreme values or outliers in the dataset.

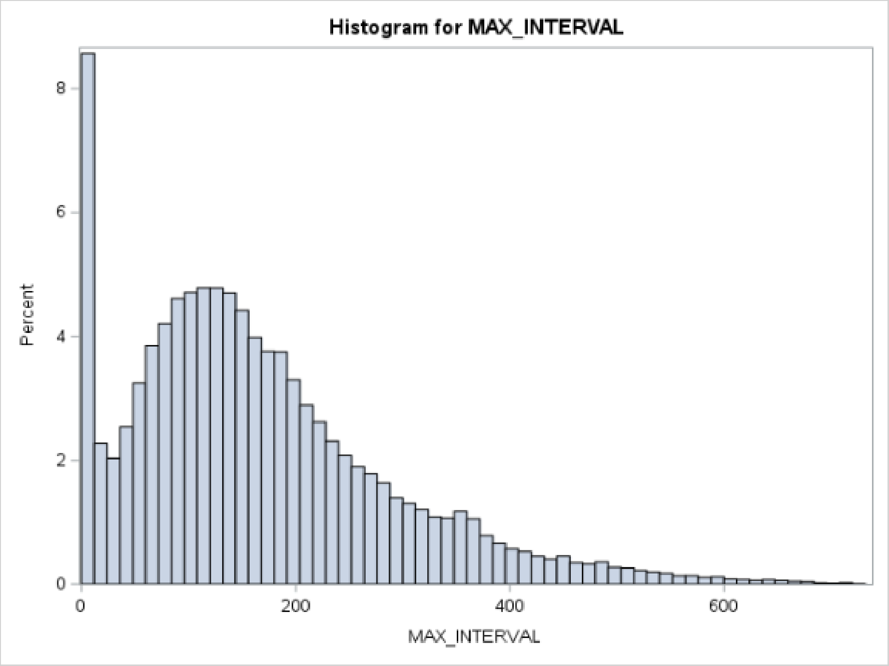


Figure 3.19.2: Histogram of MAX_INTERVAL

In this MAX_INTERVAL's histogram shows a distribution that is favourably skewed. The longest tail is formed by a gradual decline in frequency as values increase, with the highest frequency of values occurring close to zero. This long tail indicates the presence of outliers, further corroborating the finding from the box plot.

Pre-Processing Method:

Cleaning method: IQR Method for removing Outliers.

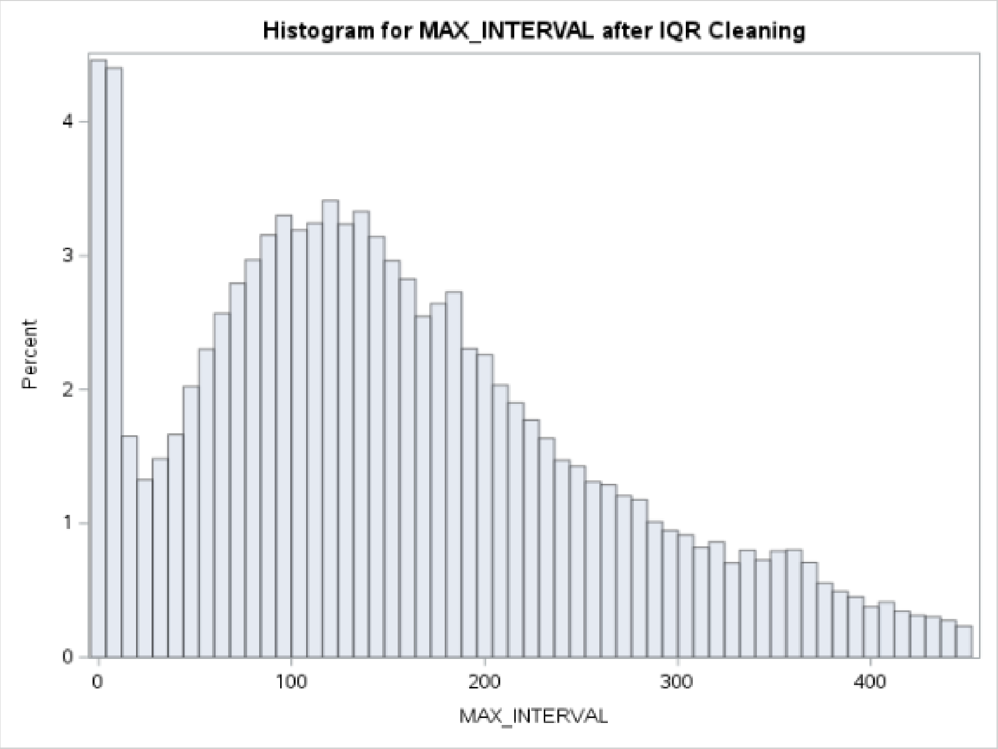


Figure 3.19.3: Histogram of MAX_INTERVAL after IQR Cleaning

To address the issue of outliers, the Interquartile Range (IQR) method was employed. This method calculates the IQR, which is the range between the first quartile (Q1) and the third quartile (Q3). Boundaries are determined beyond which data points are considered outliers, and these points are removed from the dataset.

Based on the Figure 3.19.3, the histogram after cleaning shows a more even distribution of MAX_INTERVAL values. The extreme values at the upper end have been significantly reduced, leading to a more balanced histogram. The majority of the data points are now within a more reasonable range, and the long tail is less pronounced.

After Pre-Processing:

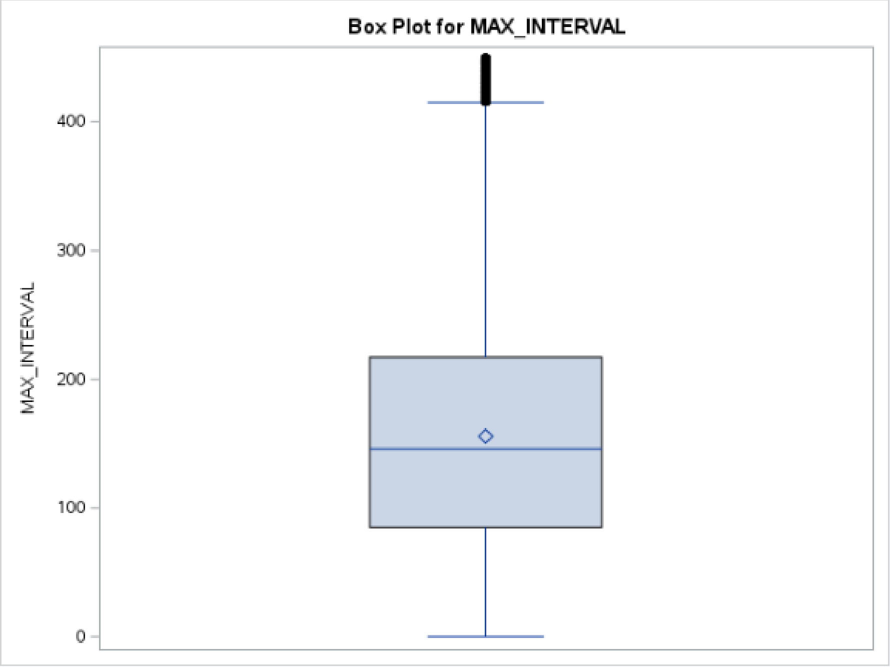


Figure 3.19.4: Updated Box Plot of MAX_INTERVAL

In this updated box plot, it demonstrates a substantial reduction in outliers. The majority of the data points now fall within the interquartile range, and the whiskers extend to a more reasonable range. This indicates a cleaner dataset with fewer extreme values. Removing these outliers provides a more accurate representation of the central tendency and variability of the MAX_INTERVAL attribute.

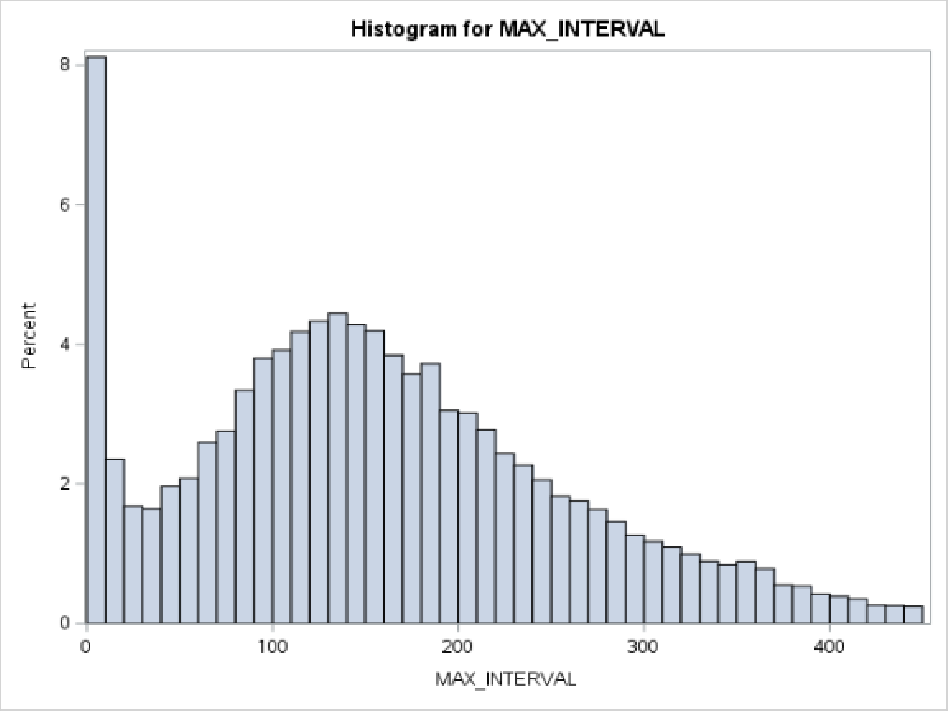


Figure 3.19.5: Updated Histogram of MAX_INTERVAL

This Histogram is the previous one for MAX_INTERVAL after cleaning. It illustrates data distribution. The overall distribution is more even and less skewed after cleaning. The frequency values decrease more steadily, and the presence of extreme values has been minimized. This results in a dataset that is more suitable for subsequent analysis, providing a clearer understanding of the MAX_INTERVAL attribute.

3.20 EXCHANGE_COUNT

Before Pre-Processing:

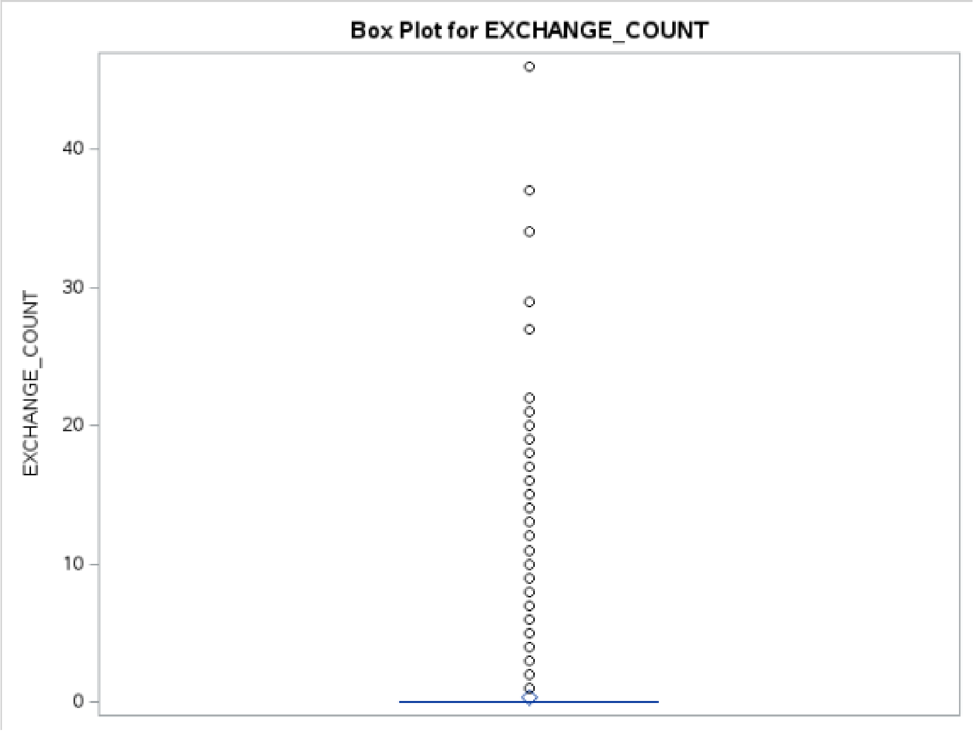


Figure 3.20.1: Box Plot of EXCHANGE_COUNT

In this box plot has showed a significant concentration of values around 0, with a several outliers until the highest value was more than 40. This indicates that most of the customers have a low exchange count, and only few of people having higher values of exchange count.

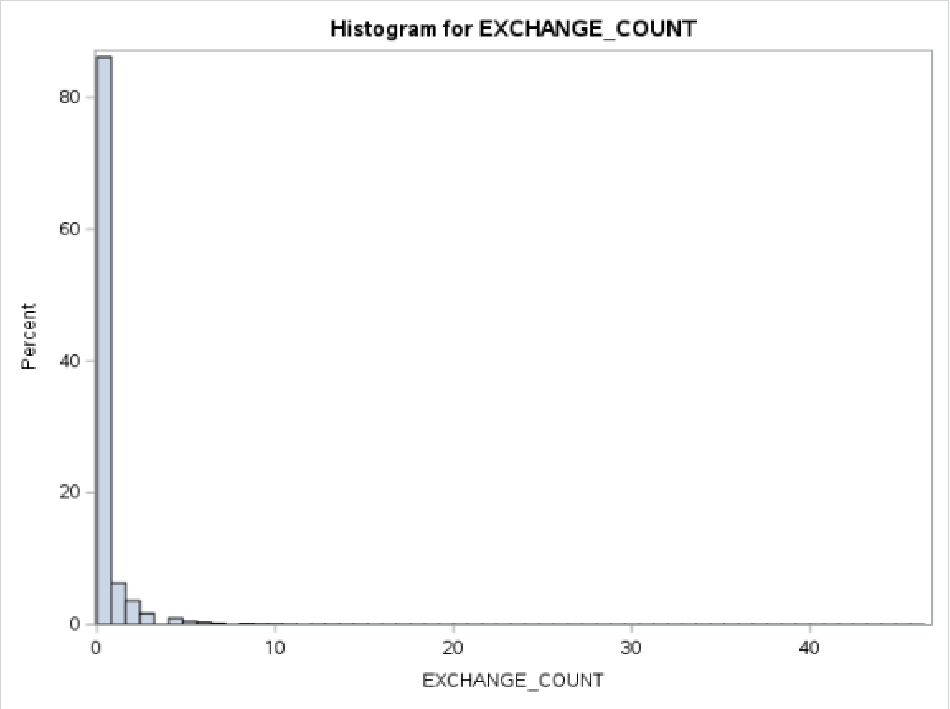


Figure 3.20.2: Histogram of EXCHANGE_COUNT

In this histogram it makes a stronger proof of confirming the box plot's finding, it shows that a skewed distribution with most values is near 0. This suggests that while most of the customers have low exchange counts, but only some have higher counts.

General Cleaning:

Cleaning Method: General Cleaning Process

Due to make the data more useful and appropriate, in general cleaning we removed the rows that have more than 2 missing values, and overall, the missing data were checked and handled.

After Pre-Processing:

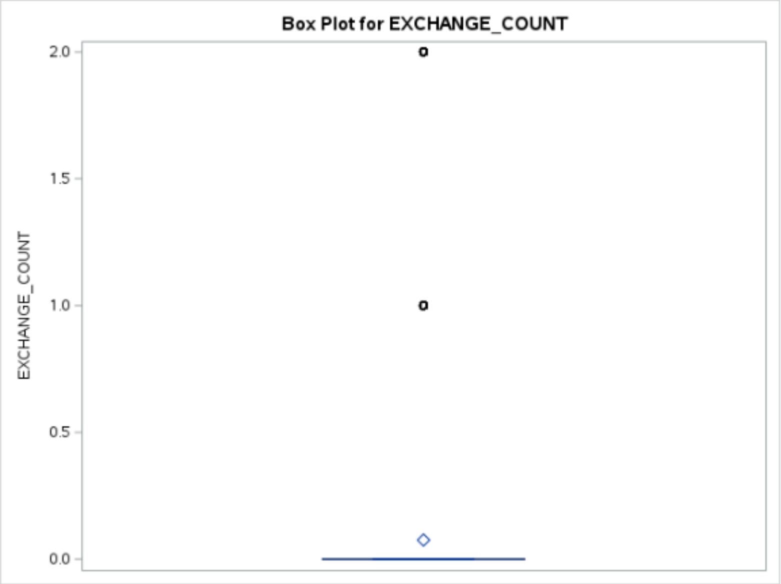


Figure 3.20.3: Updated Box Plot of EXCHANGE_COUNT

The updated box plot shows a clearer representation of the data after handling missing values and removing problematic rows.

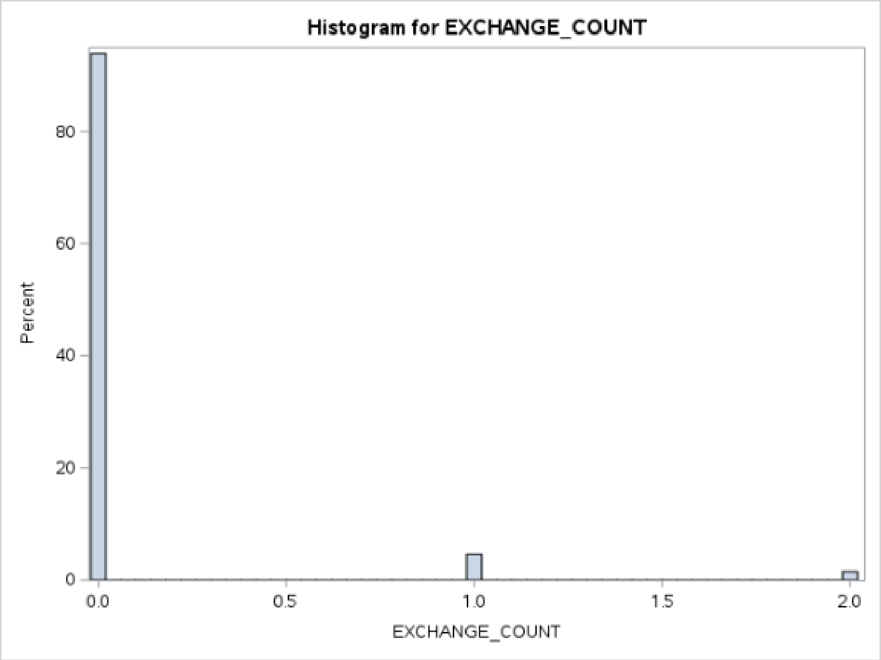


Figure 3.20.4: Updated Histogram of EXCHANGE_COUNT

The updated histogram provides a more accurate distribution of exchange counts, reflecting the changes after pre-processing.

3.21 avg_discount

Before Pre-Processing:

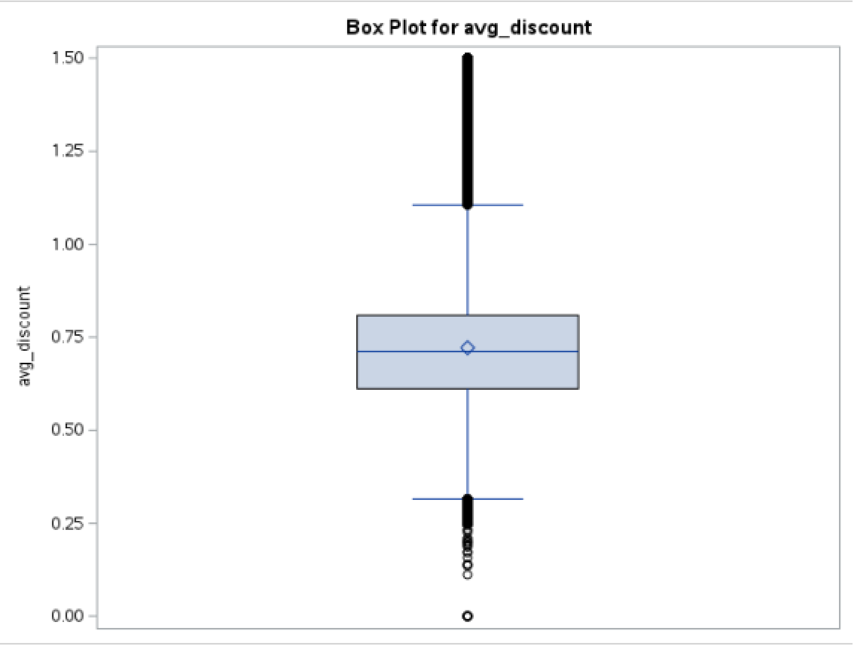


Figure 3.21.1: Box Plot of avg_discount

The box plot shows the distribution of avg_discount. The median in avg_discount is around 0.75, with a range of values and several outliers.

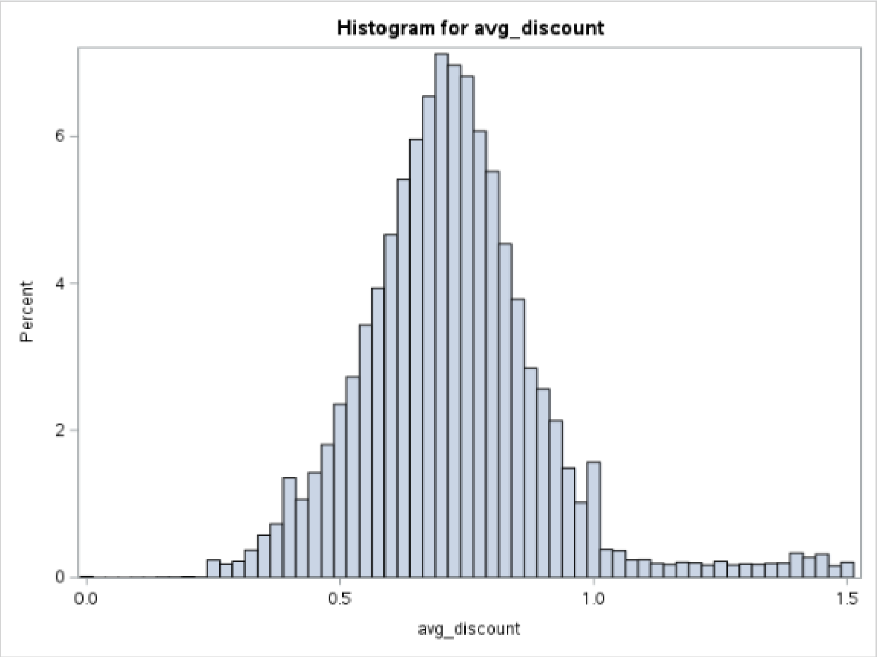


Figure 3.21.2: Histogram of avg_discount

The histogram of avg_discount, which shows in Figure 3.21.2 illustrates a roughly normal distribution centred around 0.75. There are also some extreme values exist at both of the end distribution, this indicates the variability of customer receive the discount rates.

Pre-Processing Method:

Although there are some extreme values exist in avg_discount attributes, but in this case it doesn’t affect the avg_discount variable. Therefore, there are no pre-processing was required for avg_discount.

3.22 Points_Sum

Before Pre-Processing:

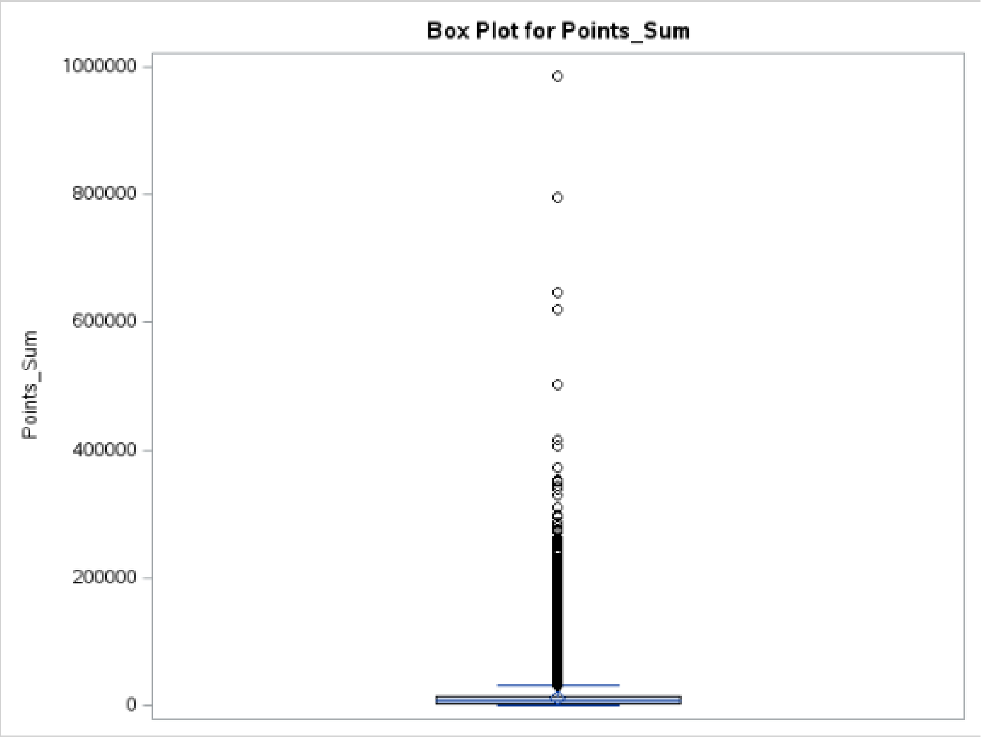


Figure 3.22.1: Box Plot of Points_Sum

The initial box plot of Points_Sum shows a significant number of outliers, with values extending well beyond the upper quartile. Besides that, there is a concentration of data points at the lower end, indicating a skewed distribution phenomenon.

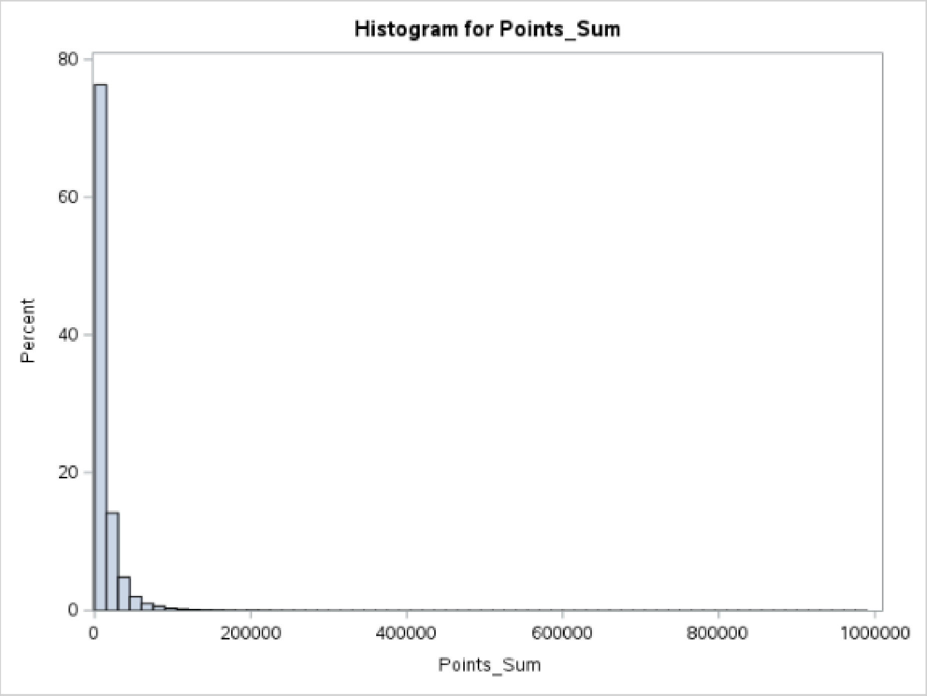


Figure 3.22.2: Histogram of Points_Sum

In this Points_Sum histogram, it corroborates the box plot by displaying a highly skewed distribution with a heavy concentration of points towards the lower end. By looking more details about it, can clearly see that a long tail is present, which indicates the presence of high-value outliers that could affect the analysis.

Pre-processing method:

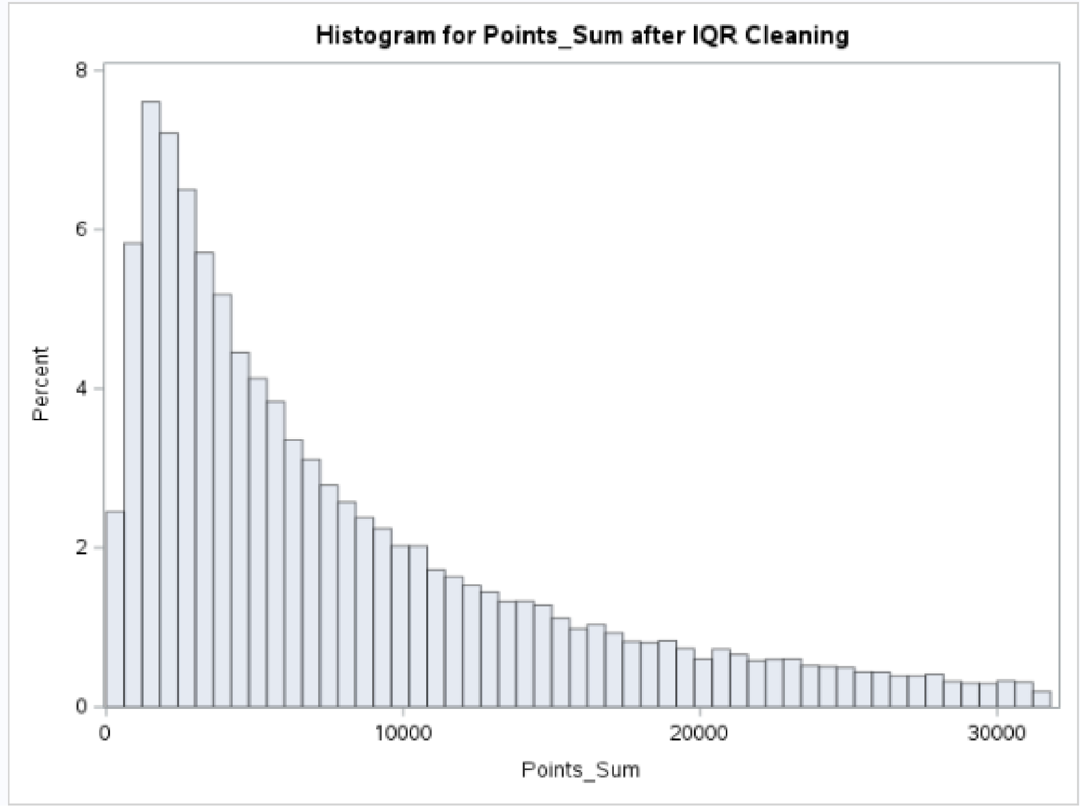


Figure 3.22.3: Histogram of Points_Sum after IQR Cleaning

Cleaning Method: IQR Method for removing Outliers.

To address the skewness and the presence of extreme outliers in Points_Sum data, we applied the Interquatile Range (IQR) method. This involves calculating the first quartile (Q1) and third quartile (Q3) and removing any data points that fall below Q1 – 1.5/QR or above Q3 + 1.5/QR. After cleaning, the histogram shows a more normalized distribution in Figure 3.22.3. The extreme values have been trimmed, reducing the skewness, and allowing for a better representation of the majority of the data points.

After Pre-Processing:

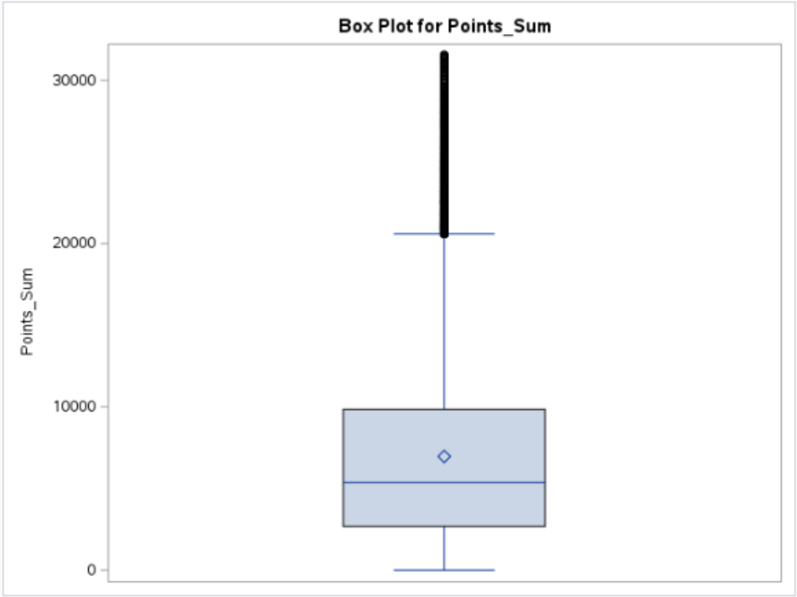


Figure 3.22.4: Updated Box Plot of Points_Sum

In this updated box plot of Points_Sum, to compare with the previous box plot, the outliers obviously become lesser, and most data points falling within a reasonable range. The upper whisker is also had been shortened significantly, indicating the successful of extreme values.

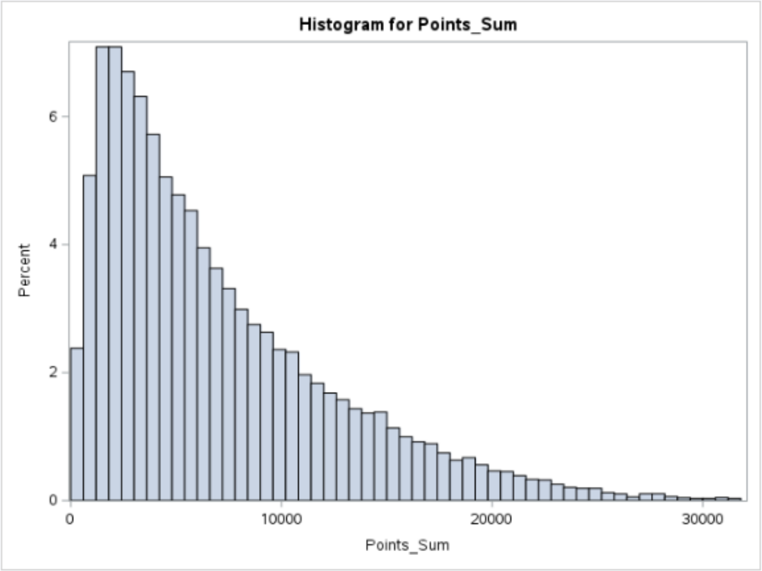


Figure 3.22.5: Updated Histogram of Points_Sum

In this updated histogram, it reflects a more balanced distribution, with the bulk of the data points clustered around the lower values but without the extreme skewness seen earlier. The distribution is now more interpretable and suitable for further statistical analysis.

3.23 Point_NotFlight

Before Pre-Processing:

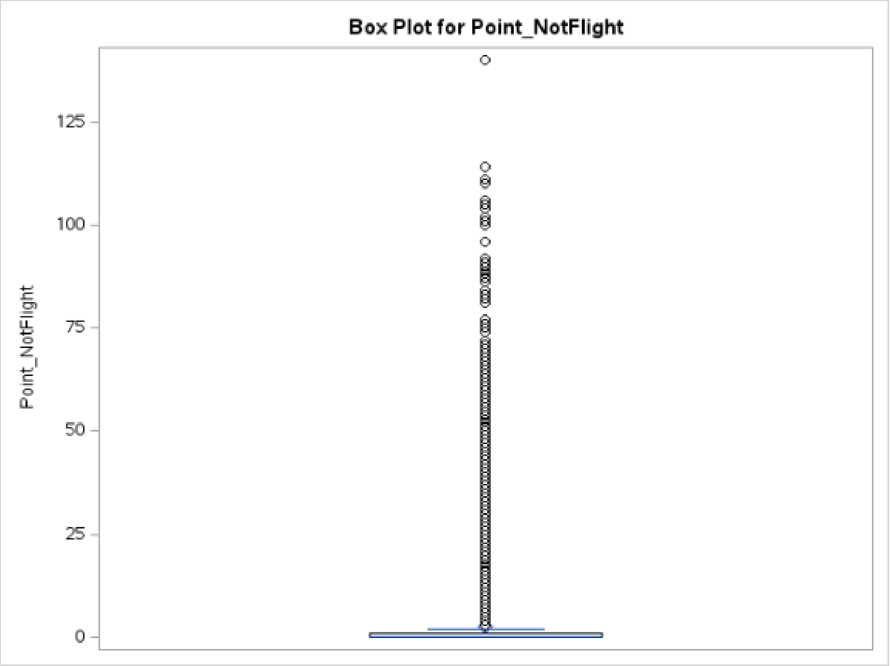


Figure 3.23.1: Box Plot of Points_NotFlight

In this box plot, it shows the Points_NotFlight's significant number of outliers with values extending well beyond the upper whisker.

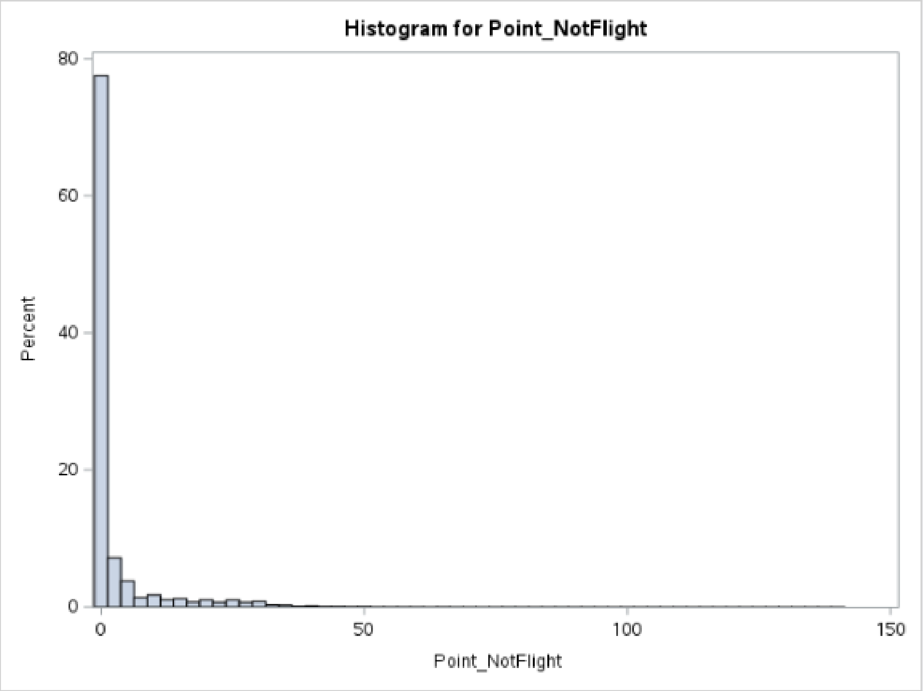
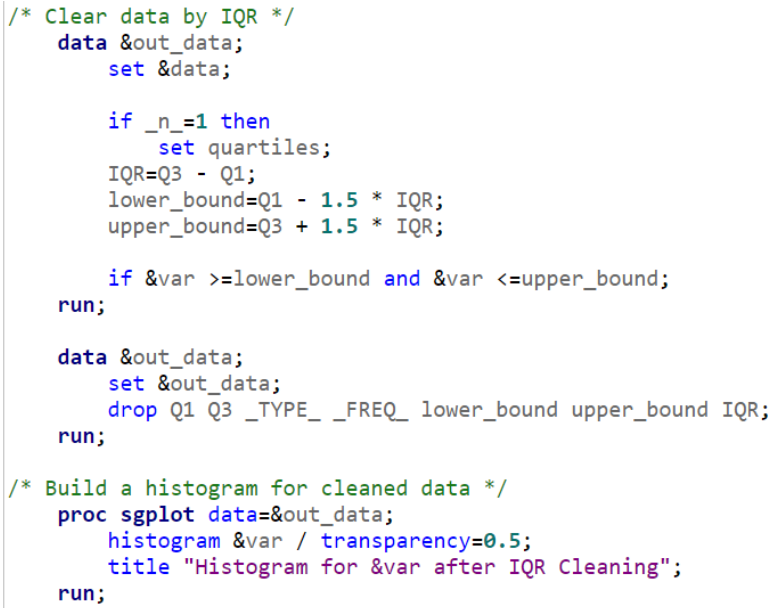


Figure 3.23.2: Histogram of Points_NotFlight

The histogram indicates a highly skewed distribution with most data points concentrated at zero and a long tail extending to higher values.

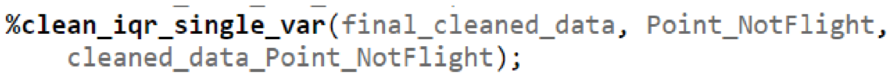
Pre-Processing Method:

Cleaning Method: IQR Method for removing Outliers.



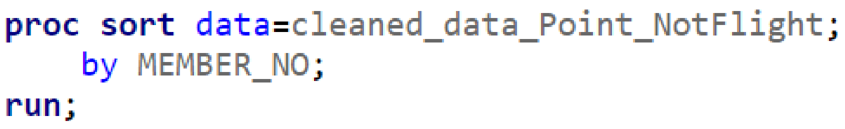
```
/* Clear data by IQR */
    data &out_data;
        set &data;

        if _n_=1 then
            set quartiles;
        IQR=Q3 - Q1;
        lower_bound=Q1 - 1.5 * IQR;
        upper_bound=Q3 + 1.5 * IQR;

        if &var >=lower_bound and &var <=upper_bound;
    run;

    data &out_data;
        set &out_data;
        drop Q1 Q3 _TYPE_ _FREQ_ lower_bound upper_bound IQR;
    run;

/* Build a histogram for cleaned data */
    proc sgplot data=&out_data;
        histogram &var / transparency=0.5;
        title "Histogram for &var after IQR Cleaning";
    run;
```

Figure 2.23.3: Code of removing outliers



```
%clean_iqr_single_var(final_cleaned_data, Point_NotFlight,
    cleaned_data_Point_NotFlight);
```

Figure 2.23.4: Code of creating histogram including Points_NotFlight



```
proc sort data=cleaned_data_Point_NotFlight;
    by MEMBER_NO;
run;
```

Figure 2.23.5: Illustrates cleaned histogram of Points_NotFlight

To address outliers, we applied the Interquartile Range (IQR) method, a widely recognized technique for outlier detection and removal. The IQR method involves calculating the first quartile (Q1) and the third quartile (Q3) of the data and defining the IQR as the difference between Q3 and Q1. Outliers are then identified as data points that fall below Q1−1.5×IQR or above Q3+1.5×IQR.

In our SAS implementation, we began by calculating Q1 and Q3 for the "Point_NotFlight" attribute. The IQR was then determined by subtracting Q1 from Q3. Next, we calculated the lower and upper bounds for outlier detection using the formulas Q1−1.5×IQR and Q3+1.5×IQR, respectively. Any data points outside these bounds were considered outliers and subsequently removed from the dataset.

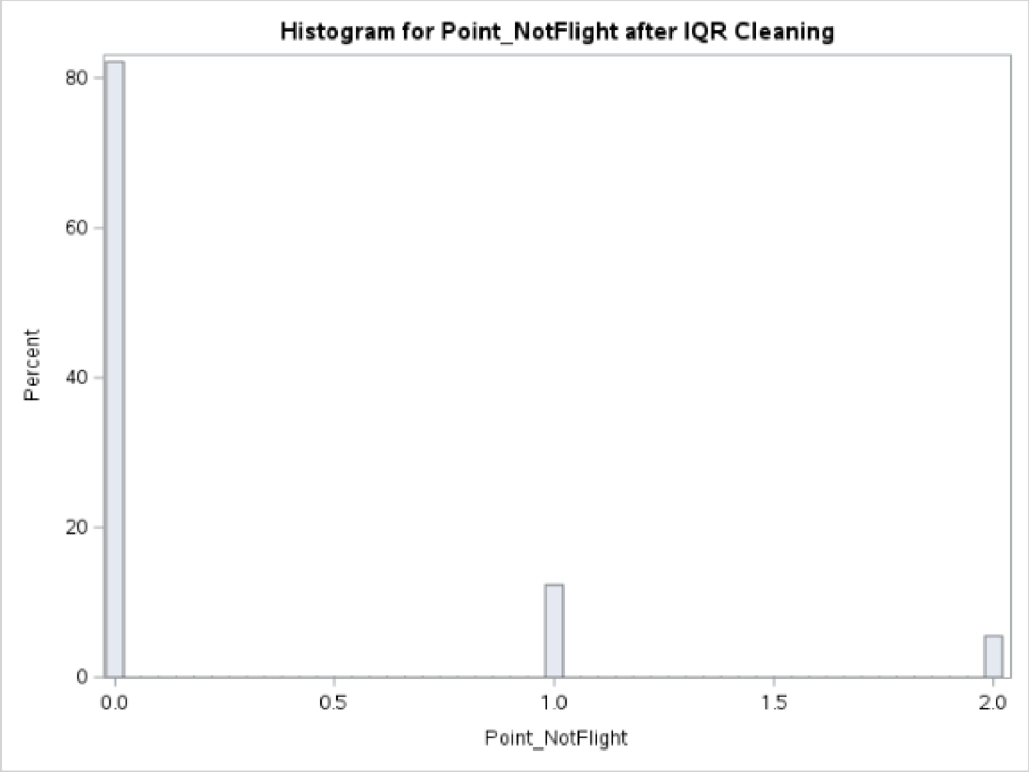


Figure 3.23.6: Histogram of Point_NotFlight after IQR Cleaning

In the Figure 3.23.6, can clearly see that this is the histogram which after applying the IQR method shows a balanced distribution. The extreme values have been removed, resulting in a dataset that is more suitable for further analysis. Here is some explanation of the cleaned histogram:

1 Concentration at Zero: Most of the data points are at zero, indicating that the majority of users have no non-flight points.

2 Reduced Spread of Values: The histogram shows a constrained range compared to the previous one, indicating the removal of extreme values, which was the outliers.

3 Presence of Small Non-Zero Values: There are a few of non-zero values at 1.0 and 2.0, indicate some of the users accumulate a small number of non-flight points.

4 Improved Data Quality: After remove all of the outliers, the histogram results a more representative and meaningful distribution of the data, improving the reliability for further analysis.
5 Data Interpretation: While most of the users do not accumulate non-flight points, there are occasional small accumulations, reflecting more accurate user behaviour.

After Pre-Processing:

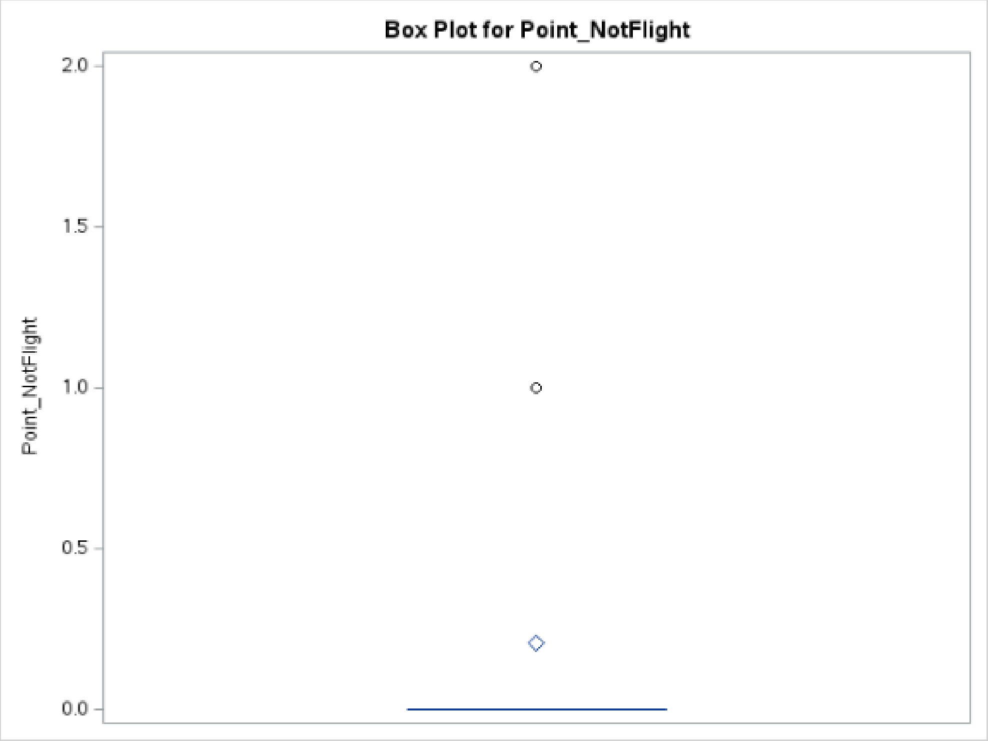


Figure 3.23.7: Updated Box Plot of Points_NotFlight

In this Figure 3.23.7, the updated box plot demonstrates a significant reduction in outliers. After using IQR method, it shows that there are lots of the extreme values have been removed in this previous box plot. There are still have a few outliers existing, but significantly is lesser than before, and makes the dataset be more reliable than the previous one.

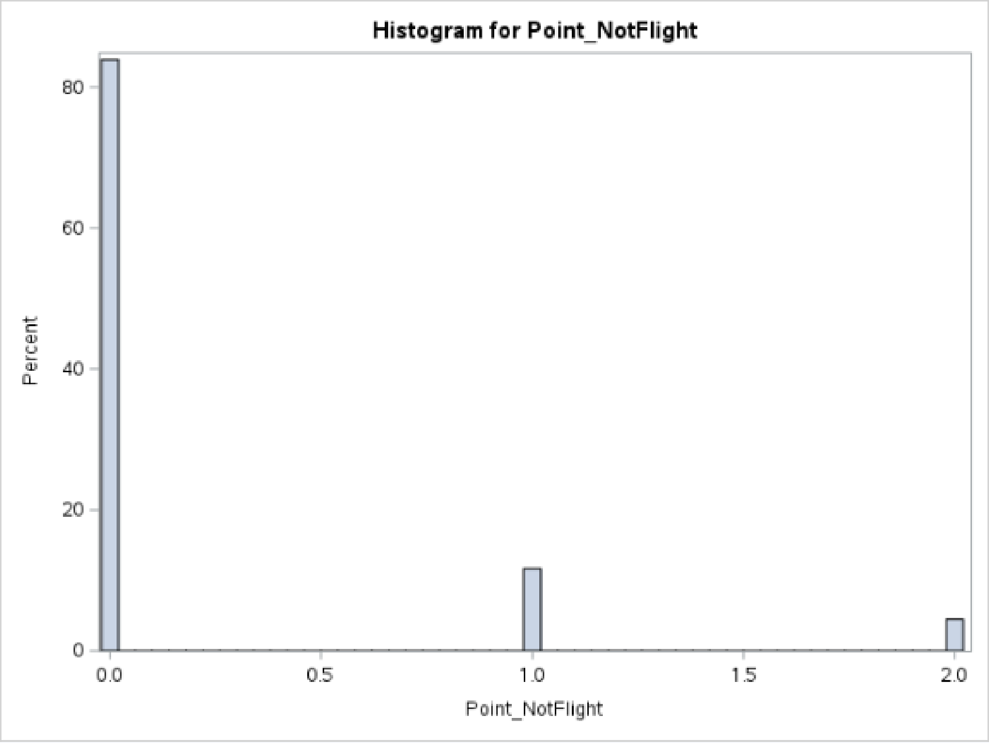


Figure 3.23.8: Updated Histogram of Points_NotFlight

In this updated histogram of Points_NotFlight a large portion of data points are at zero, indicating that the majority of users have no non-flight points. The histogram shows a small peak at 1.0 and 2.0, proven that some users accumulate a small number of non-flight points. These values are now more discernible and less overshadowed by extreme values. By improving the data quality, it allows for a better insight and more accurate analysis by reflecting the true user behaviour without the distortion caused by outliers. Therefore, improve the data quality could enhanced the reliability of the data and make more accurate decision making.

3.24 Overall

Before Pre-Processing:

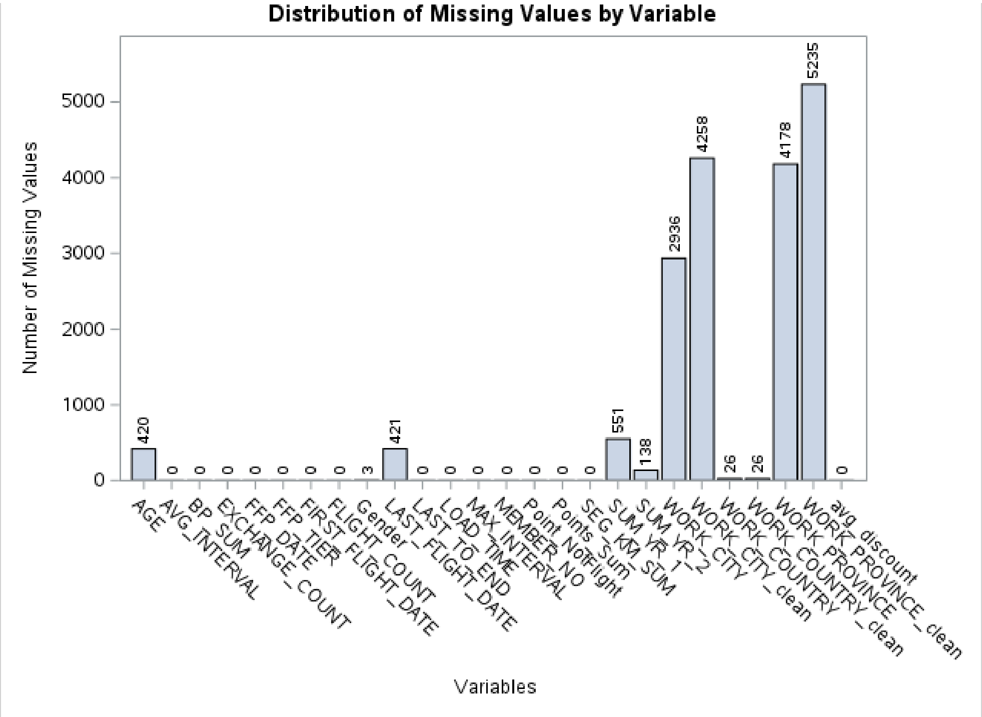
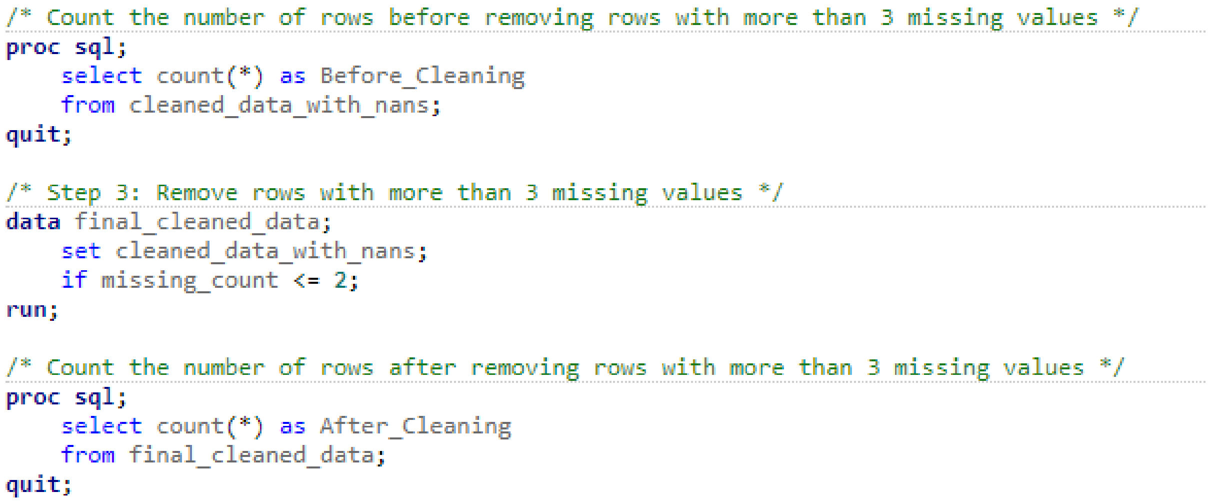


Figure 3.24.1: Distribution of Missing Values by Variable

This bar chart illustrates the distribution of missing values across various variables in the dataset. The x-axis represents the number of missing values, while the y-axis shows the frequency of variables with those missing values.

Pre-Processing Method:

Cleaning Method: Remove rows that have more than 2 missing values.



```
/* Count the number of rows before removing rows with more than 3 missing values */
proc sql;
    select count(*) as Before_Cleaning
    from cleaned_data_with_nans;
quit;

/* Step 3: Remove rows with more than 3 missing values */
data final_cleaned_data;
    set cleaned_data_with_nans;
    if missing_count <= 2;
run;

/* Count the number of rows after removing rows with more than 3 missing values */
proc sql;
    select count(*) as After_Cleaning
    from final_cleaned_data;
quit;
```

Figure 3.24.3: Code of removing rows more than 2 missing values

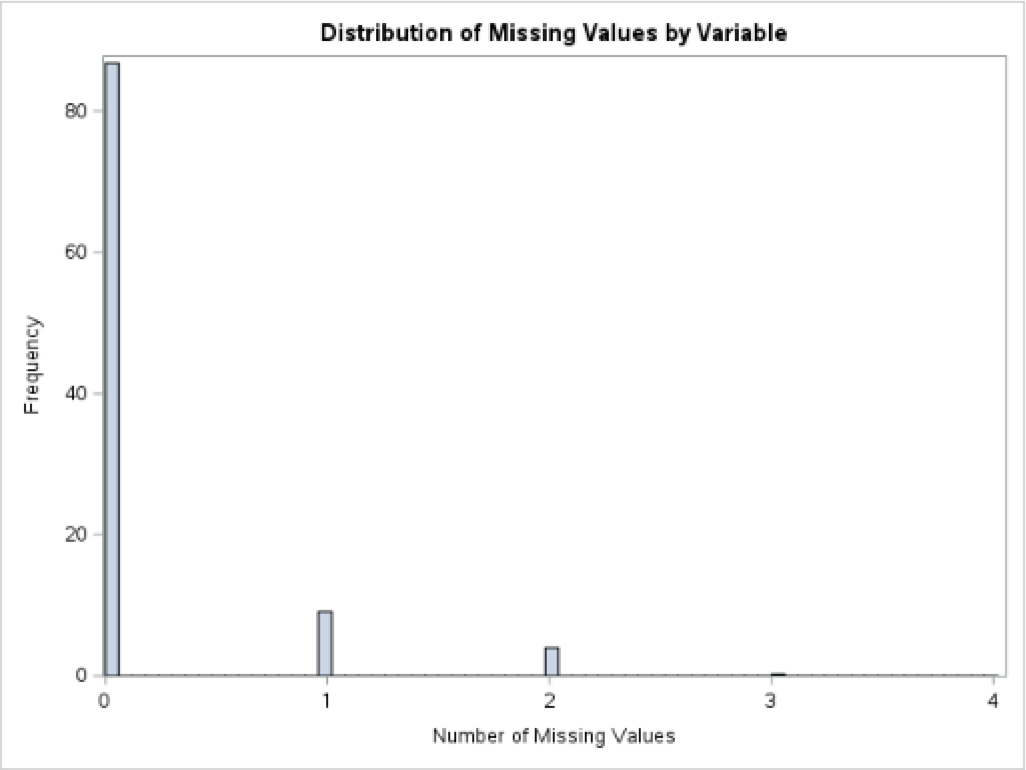
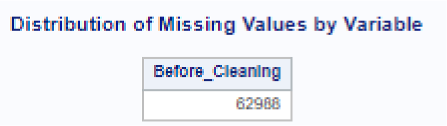


Figure 3.24.4: The frequency of missing values by Variable

The majority of the variables have zero missing values, as indicated by the highest bar at zero on the x-axis. A smaller number of variables have one or two missing values, shown by the shorter bars at one and two on the x-axis. Very few variables have three or more missing values, indicating that missing data is relatively uncommon in this dataset.

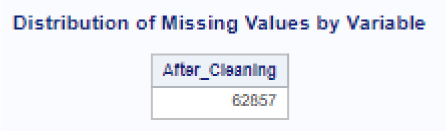


Distribution of Missing Values by Variable

Before_Cleaning
62988

Figure 3.24.5

The dataset initially contained a total of 62,988 entries with varying numbers of missing values across different variables.



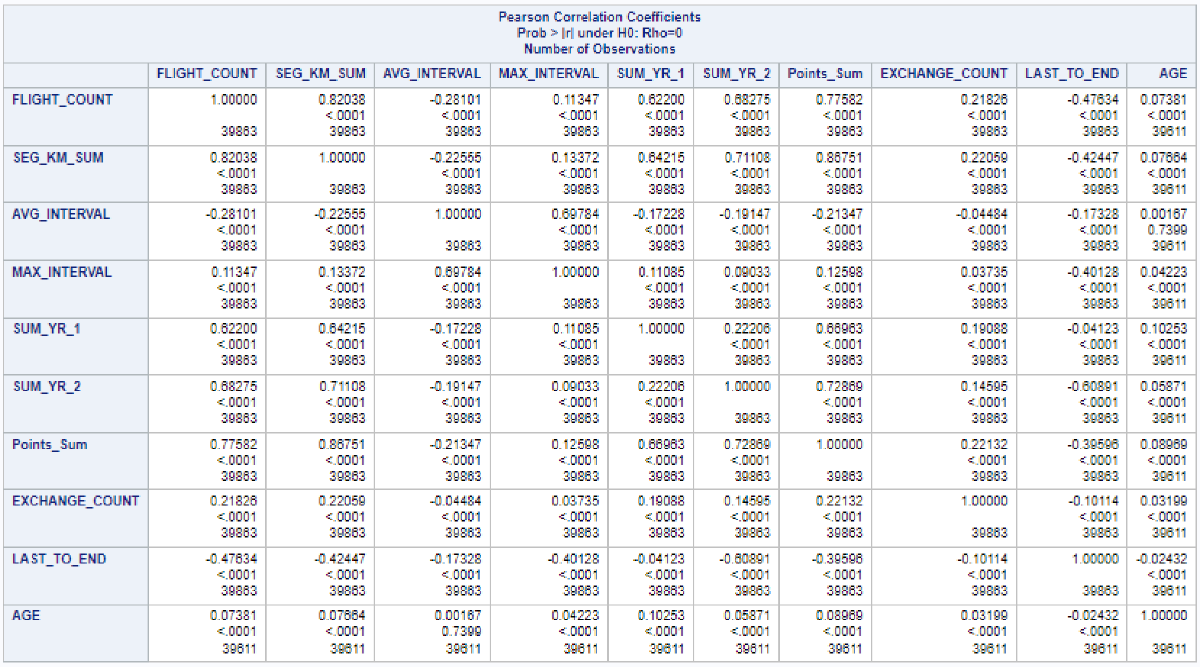
Distribution of Missing Values by Variable

After_Cleaning
62857

Figure 3.24.6

Post-cleaning, the dataset was reduced to 62,857 entries. This reduction indicates that rows with excessive missing values (more than two) were removed to improve data integrity.

Cleaning Method: Data Correlation



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	FLIGHT_COUNT	SEG_KM_SUM	AVG_INTERVAL	MAX_INTERVAL	SUM_YR_1	SUM_YR_2	Points_Sum	EXCHANGE_COUNT	LAST_TO_END	AGE
FLIGHT_COUNT	1.00000 39863	0.82038 <.0001 39863	-0.28101 <.0001 39863	0.11347 <.0001 39863	0.62200 <.0001 39863	0.68275 <.0001 39863	0.77582 <.0001 39863	0.21826 <.0001 39863	-0.47634 <.0001 39863	0.07381 <.0001 39611
SEG_KM_SUM	0.82038 <.0001 39863	1.00000 39863	-0.22555 <.0001 39863	0.13372 <.0001 39863	0.64215 <.0001 39863	0.71108 <.0001 39863	0.86751 <.0001 39863	0.22059 <.0001 39863	-0.42447 <.0001 39863	0.07664 <.0001 39611
AVG_INTERVAL	-0.28101 <.0001 39863	-0.22555 <.0001 39863	1.00000 39863	0.69784 <.0001 39863	-0.17228 <.0001 39863	-0.19147 <.0001 39863	-0.21347 <.0001 39863	-0.04484 <.0001 39863	-0.17328 <.0001 39863	0.00167 0.7399 39611
MAX_INTERVAL	0.11347 <.0001 39863	0.13372 <.0001 39863	0.69784 <.0001 39863	1.00000 39863	0.11085 <.0001 39863	0.09033 <.0001 39863	0.12598 <.0001 39863	0.03735 <.0001 39863	-0.40128 <.0001 39863	0.04223 <.0001 39611
SUM_YR_1	0.62200 <.0001 39863	0.64215 <.0001 39863	-0.17228 <.0001 39863	0.11085 <.0001 39863	1.00000 39863	0.22206 <.0001 39863	0.66963 <.0001 39863	0.19088 <.0001 39863	-0.04123 <.0001 39863	0.10253 <.0001 39611
SUM_YR_2	0.68275 <.0001 39863	0.71108 <.0001 39863	-0.19147 <.0001 39863	0.09033 <.0001 39863	0.22206 <.0001 39863	1.00000 39863	0.72869 <.0001 39863	0.14595 <.0001 39863	-0.60891 <.0001 39863	0.05871 <.0001 39611
Points_Sum	0.77582 <.0001 39863	0.86751 <.0001 39863	-0.21347 <.0001 39863	0.12598 <.0001 39863	0.66963 <.0001 39863	0.72869 <.0001 39863	1.00000 39863	0.22132 <.0001 39863	-0.39596 <.0001 39863	0.08969 <.0001 39611
EXCHANGE_COUNT	0.21826 <.0001 39863	0.22059 <.0001 39863	-0.04484 <.0001 39863	0.03735 <.0001 39863	0.19088 <.0001 39863	0.14595 <.0001 39863	0.22132 <.0001 39863	1.00000 39863	-0.10114 <.0001 39863	0.03199 <.0001 39611
LAST_TO_END	-0.47634 <.0001 39863	-0.42447 <.0001 39863	-0.17328 <.0001 39863	-0.40128 <.0001 39863	-0.04123 <.0001 39863	-0.60891 <.0001 39863	-0.39596 <.0001 39863	-0.10114 <.0001 39863	1.00000 39863	-0.02432 <.0001 39611
AGE	0.07381 <.0001 39611	0.07664 <.0001 39611	0.00167 0.7399 39611	0.04223 <.0001 39611	0.10253 <.0001 39611	0.05871 <.0001 39611	0.08969 <.0001 39611	0.03199 <.0001 39611	-0.02432 <.0001 39611	1.00000 39611

Figure 3.24.7: Pearson Correlation Coefficients

The Pearson correlation coefficients table highlights the need for logarithmic transformations for variables with high correlations (above 0.6). By applying these transformations, we can enhance the stability and normality of the data, leading to more reliable and interpretable statistical analyses. The transformed data will be better suited for regression models and other analytical techniques, improving the accuracy and robustness of the findings.

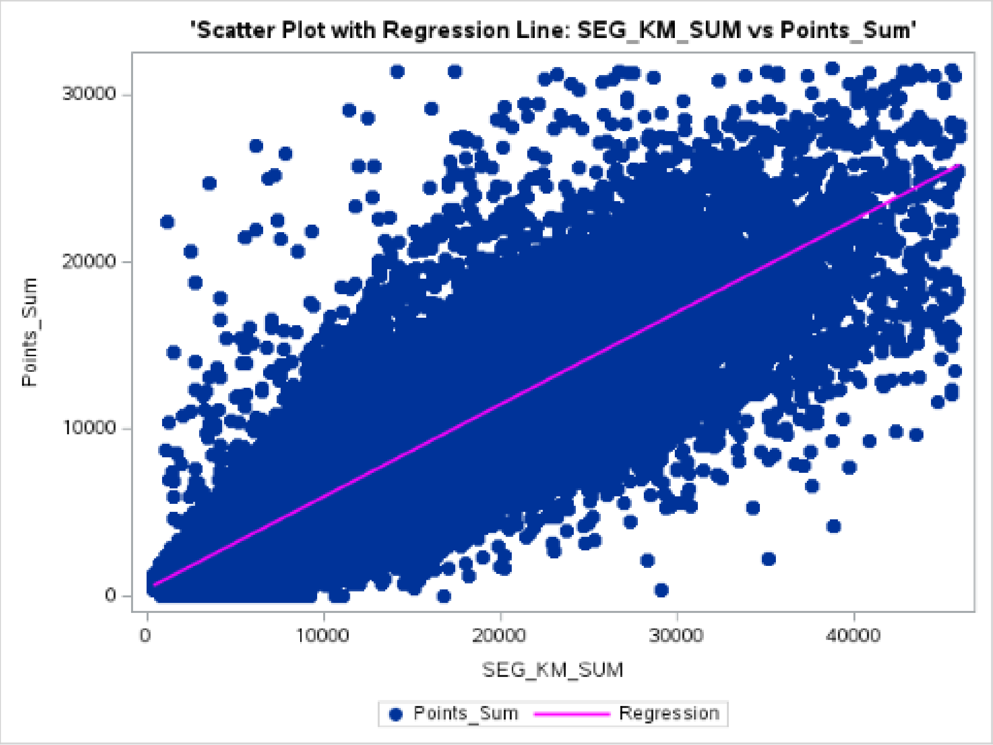


Figure 3.24.8: Scatter Plot with regression Line: SEG_KM_SUM vs Points_Sum

In data analysis, understanding the correlation between variables is crucial. Correlation measures the strength and direction of the relationship between two variables. A positive correlation occurs when an increase in one variable is associated with an increase in another variable, which is depicted by an upward sloping line on a scatter plot. Conversely, a negative correlation occurs when an increase in one variable is associated with a decrease in another variable, shown by a downward sloping line. The provided scatter plot illustrates a positive correlation between SEG_KM_SUM and Points_Sum, as evidenced by the upward sloping regression line.

To identify columns with high correlation, we examine the correlation coefficients. High correlation is indicated by a coefficient value close to 1 (positive correlation) or -1 (negative correlation). In our analysis, the columns SEG_KM_SUM and Points_Sum show a strong positive correlation.

Correlation graphs are useful for visually representing the relationships between pairs of variables. These graphs help in identifying trends and dependencies. For our analysis, we will focus on graphs that clearly demonstrate these dependencies. Log transformation is a technique used to linearize relationships between variables, especially when the data spans several orders of magnitude. By applying log transformation to the variables identified from the correlation graphs, we can often achieve a more linear relationship, making it easier to analyse and interpret the data.

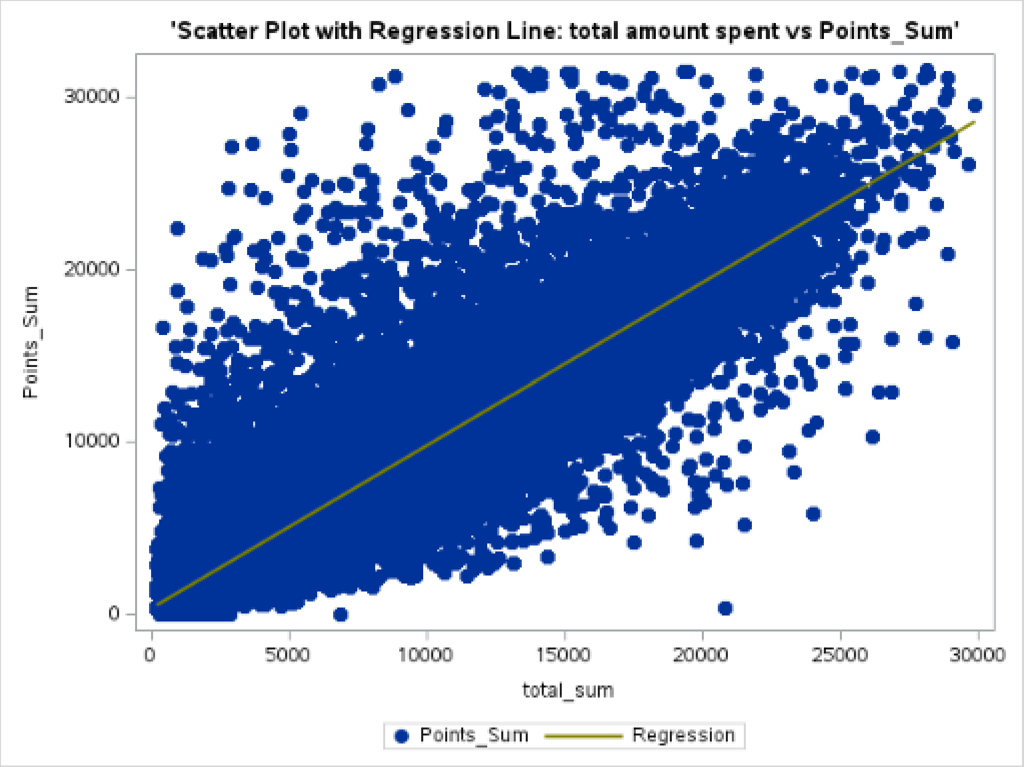


Figure 3.24.9: Scatter Plot with Regression Line: total amount spent vs Points_Sum

By analysing the correlation between total_sum and Points_Sum, we identified a significant positive correlation, indicating that as one variable increases, so does the other. This relationship was critical for understanding how spending behaviour affects points accumulation.

Correlation analysis helped validate the integrity of the data. High correlation values between expected related attributes, such as total_sum and Points_Sum, confirmed that the data was consistent and logical, with no major anomalies or discrepancies that would undermine the analysis.

Understanding the correlation between attributes informed our data cleaning strategy. For example, ensuring that outliers did not disproportionately affect the regression model was crucial. Identifying highly correlated variables allowed us to focus our cleaning efforts on ensuring these relationships remained intact and meaningful.

By maintaining and confirming the strong correlation between these variables, we improved the predictive accuracy of the regression model. The scatter plot and regression line serve as visual confirmations of the underlying data quality and the effectiveness of our pre-processing methods.

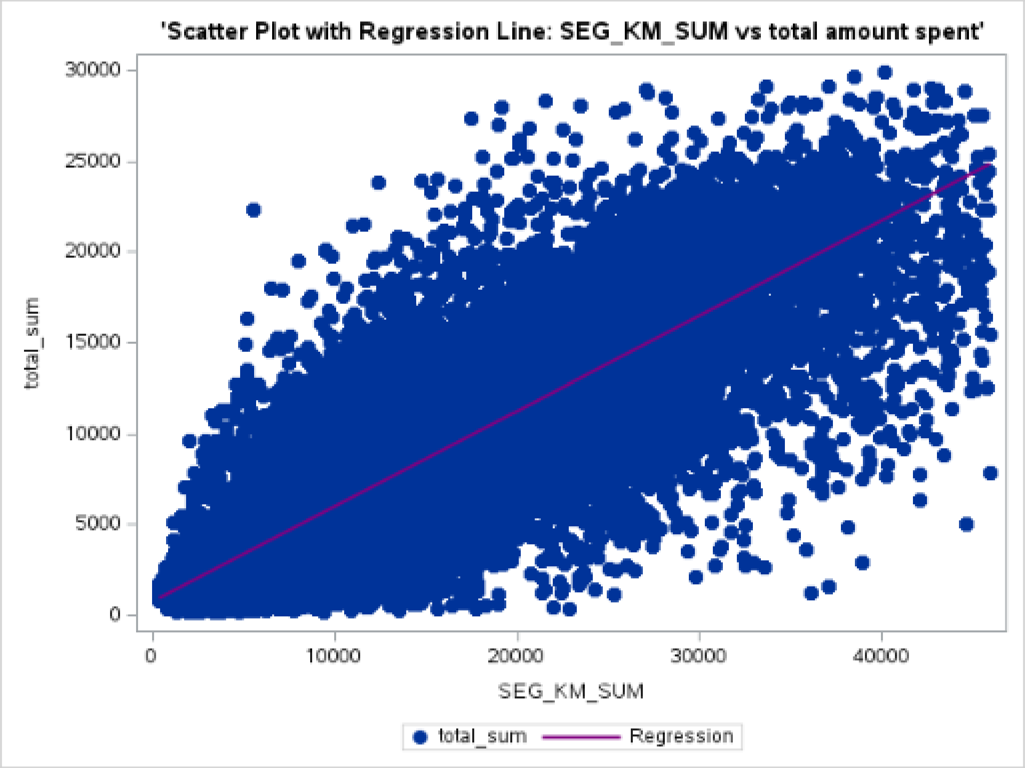


Figure 3.24.10: The Scatter Plot with Regression Line: SEG_KM_SUM vs total amount spent.

In Figure 3.24.10, we examine the relationship between the total kilometres flown by customers, which was SEG_KM_SUM and the total amount spent. The scatter plot is accompanied by a regression line that visually represents the trend within the data. The primary observation from this figure is the positive correlation between SEG_KM_SUM and the total amount spent. As the total kilometres flown by customers increase, so does their spending, which is depicted by the upward-sloping regression line.

The data points in the scatter plot are widely spread, indicating a significant variability in spending among customers with similar flight distances. Despite this spread, the regression line provides a clear indication of the overall trend. This positive relationship suggests that more frequent flyers or those who travel longer distances tend to spend more. This insight is valuable for customer segmentation, allowing businesses to identify high-value customers who may be more responsive to targeted marketing strategies and loyalty programs. Understanding this relationship helps in designing promotional offers and rewards that align with the spending patterns of frequent travellers, ultimately enhancing customer satisfaction and retention.

Cleaning Method: Data Logarithms (Data Transformation)

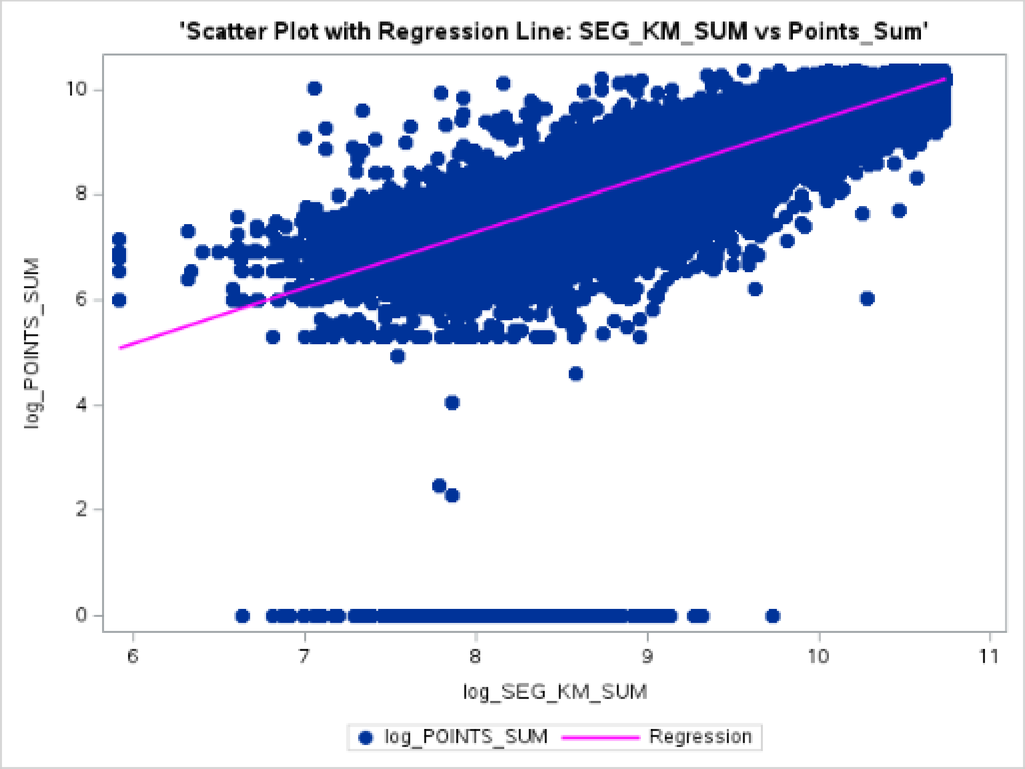


Figure 3.24.11: Scatter plot with Regression Line: SEG_KM_SUM vs Points_Sum

Logarithmic transformation is applied to linearize the data, particularly when the original relationship between variables such as SEG_KM_SUM and POINTS_SUM is exponential or follows a power-law distribution. By transforming the data logarithmically, the relationship becomes more linear, allowing for a better fit with a linear regression model. Additionally, if the original data for SEG_KM_SUM and POINTS_SUM are highly skewed, the logarithmic transformation can help in normalizing the distribution, thereby improving the regression analysis. This transformation also addresses heteroscedasticity by stabilizing the variance across the range of predictor values, ensuring that the variability of the response variable is consistent. Moreover, the results from a regression of log-transformed variables are more interpretable, as the coefficients can be understood in terms of percentage changes. This interpretability is particularly advantageous in contexts like economics and finance, where understanding proportional changes is often more intuitive.

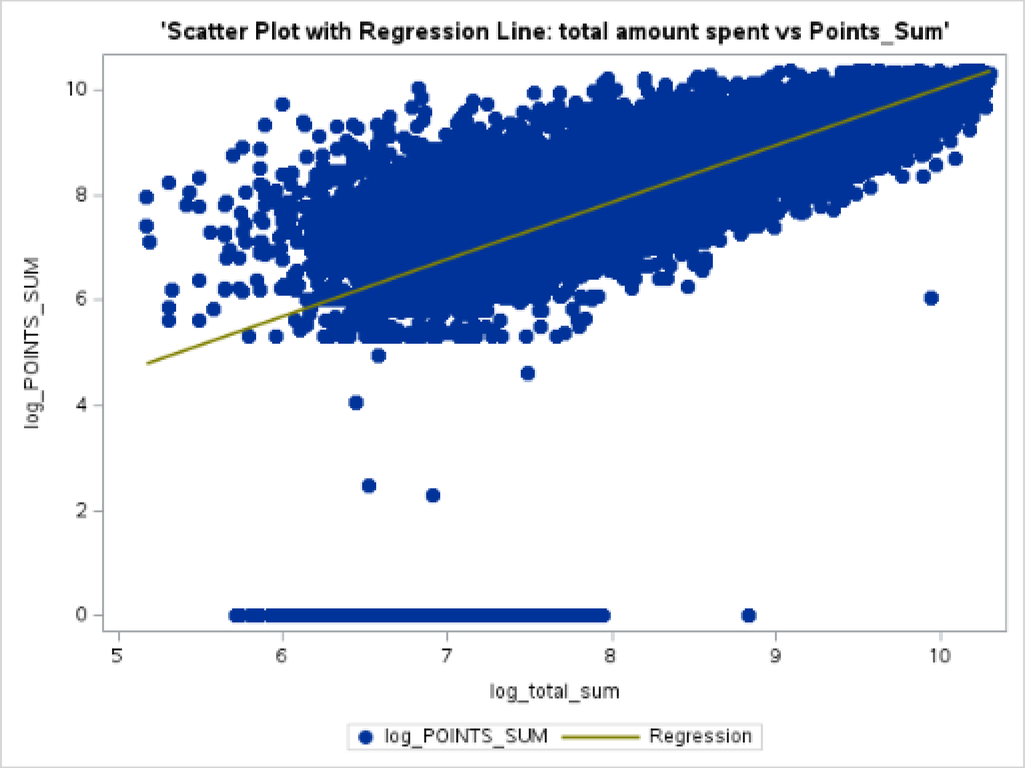


Figure 3.24.12: Scatter Plot with Regression Line: total amount spent vs Points_Sum

This scatter plot illustrates the relationship between the log-transformed total amount spent (log_total_sum) and the log-transformed cumulative points (log_POINTS_SUM). Each blue dot represents an individual data point, and the green line is the regression line fitted to the log-transformed data. There is a positive correlation between the log-transformed total amount spent and the log-transformed cumulative points. As one increases, so does the other.

The regression line shows a clear, linear relationship between the log-transformed variables, indicating a consistent trend. The data points are more uniformly spread around the regression line, indicating reduced variability after log transformation.

Logarithmic transformation reduces skewness in the data, compressing the range of values and resulting in a more symmetrical distribution. The transformation stabilizes variance, making the data points' spread more uniform and enhancing the linear relationship between variables.

The transformation improves the linearity of the correlation, making the relationship between variables more interpretable and reliable. The regression model fits better to the log-transformed data, capturing the relationship more accurately and leading to better predictive performance.

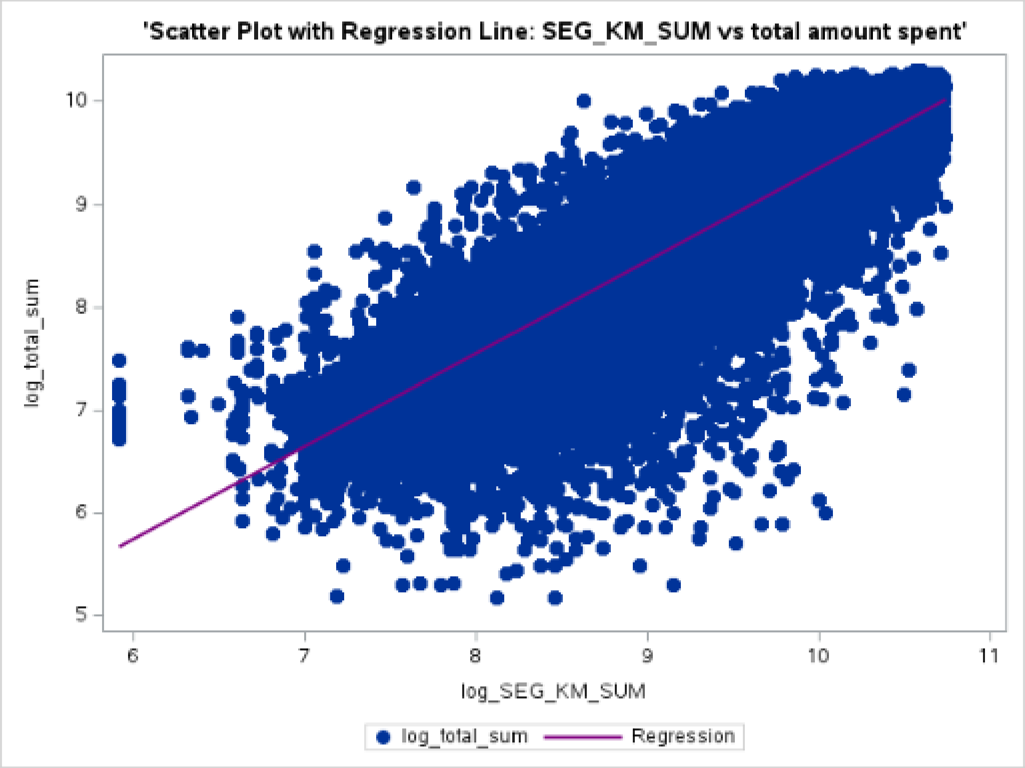


Figure 3.24.13: Data Logarithms of SEG_KM_SUM vs total amount spent.

In this Figure 3.24.13, it presents a scatter plot that explores the relationship between the logarithm of total kilometers flown log_SEG_KM_SUM and the logarithm of the total amount spent (log_total_sum). This transformation stabilizes the variance and makes the relationship between these variables more linear, as seen in the plot.

The regression line in this figure 3.24.13 also demonstrates a strong positive correlation, similar to the raw data in Figure 3.24.3. However, the logarithmic transformation reduces the skewness of the data, allowing for a clearer and more linear relationship. This transformation is particularly useful in addressing heteroscedasticity, where the variability of one variable is unequal across the range of another variable. By applying the logarithmic transformation, the scatter plot becomes more symmetrical and the variance more consistent, making it easier to interpret and model.

This figure 3.24.13 underscores the importance of data transformation in enhancing the interpretability and accuracy of linear models. The clear linear trend after the logarithmic transformation indicates that customers' spending increases proportionately with their flight distance, even when considering multiplicative effects. This insight can further refine predictive models, enabling businesses to better forecast customer spending based on their travel behaviour. By leveraging this transformed data, companies can develop more effective marketing strategies and personalized customer experiences, ultimately driving higher engagement and revenue.

Comparison Histogram:

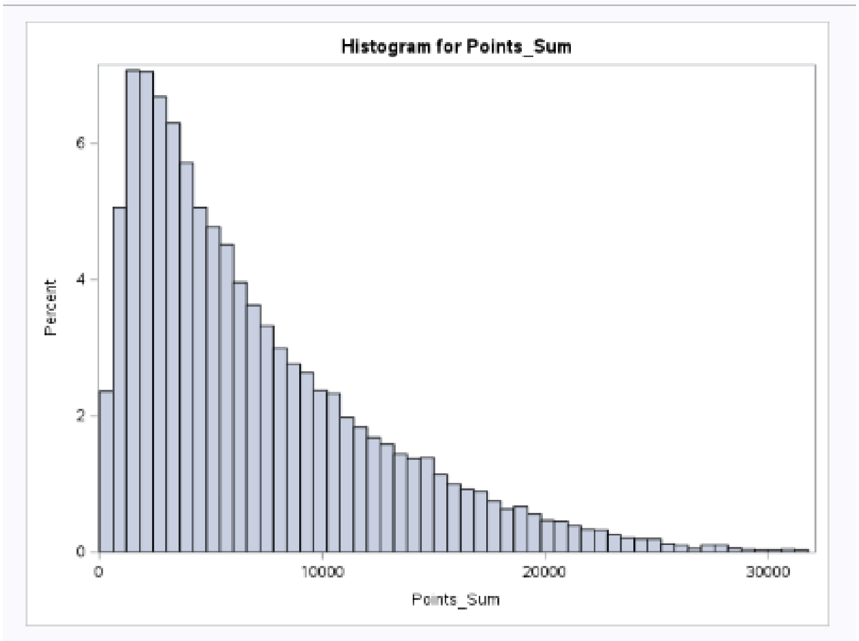


Figure 3.24.14: Before using logarithms in POINTS_SUM

Before applying the logarithmic transformation, the histogram of POINTS_SUM reveals a highly skewed distribution. The majority of the data points are concentrated towards the lower end of the spectrum, with the frequency decreasing rapidly as the values increase. This type of distribution indicates that most customers have accumulated relatively low point sums, while a smaller number of customers have significantly higher point sums. The long tail extending towards the higher values suggests the presence of outliers, which can potentially distort the analysis and interpretation.

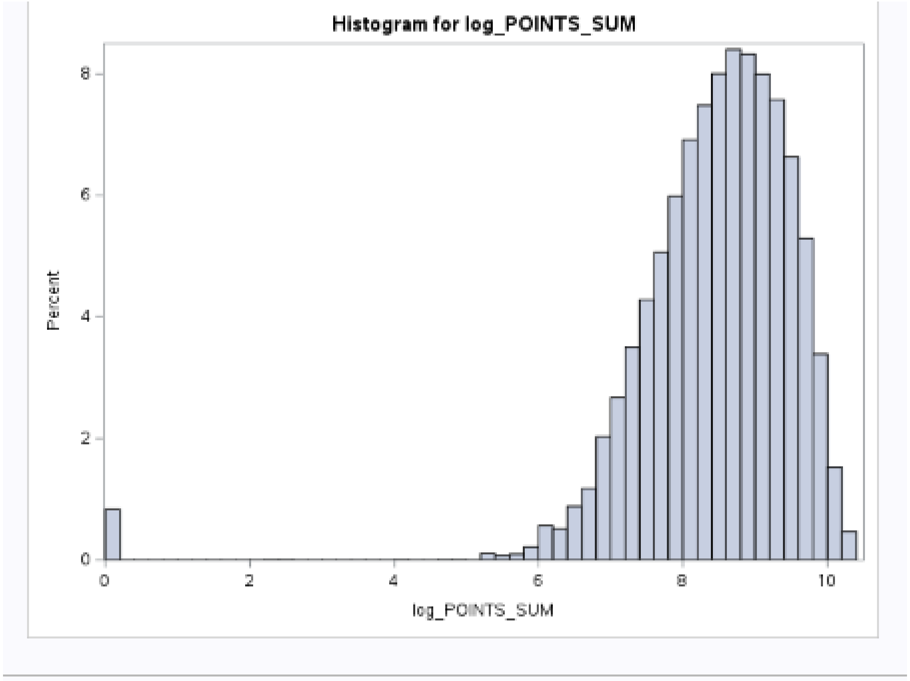


Figure 3.24.15: After using logarithms in POINTS_SUM

After using the logarithms transformation, the histogram of log_POINTS_SUM shows a more normalized and symmetric distribution. The transformation compresses the range of values, reducing the impact of extreme outliers and making the data more evenly distributed. The shape of the histogram is now more bell-shaped, resembling a normal distribution. This transformation helps in stabilizing the variance and making the data more suitable for further statistical analysis, such as regression modelling.

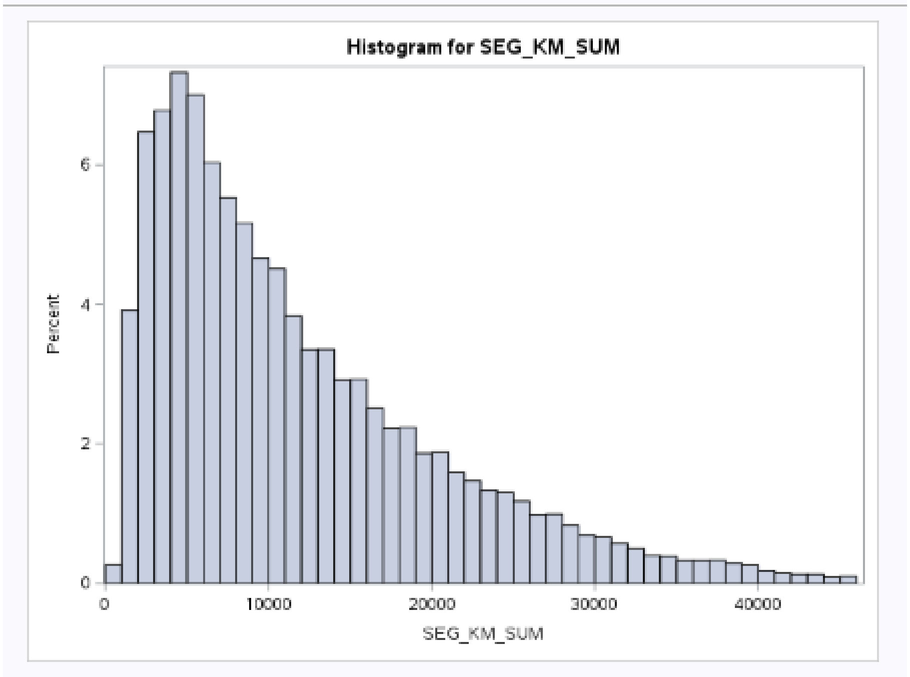


Figure 3.24.16: Before using logarithms in SEG_KM_SUM

The histogram of SEG_KM_SUM before applying a logarithmic transformation (Figure 3.24.16) illustrates a right-skewed distribution. The majority of the data points are concentrated at lower values, with a sharp decline as the values increase. This indicates that most customers have travelled relatively shorter distances, with fewer customers traveling longer distances. The skewness of this distribution can pose challenges for certain statistical analyses and machine learning models, as it violates the assumption of normality.

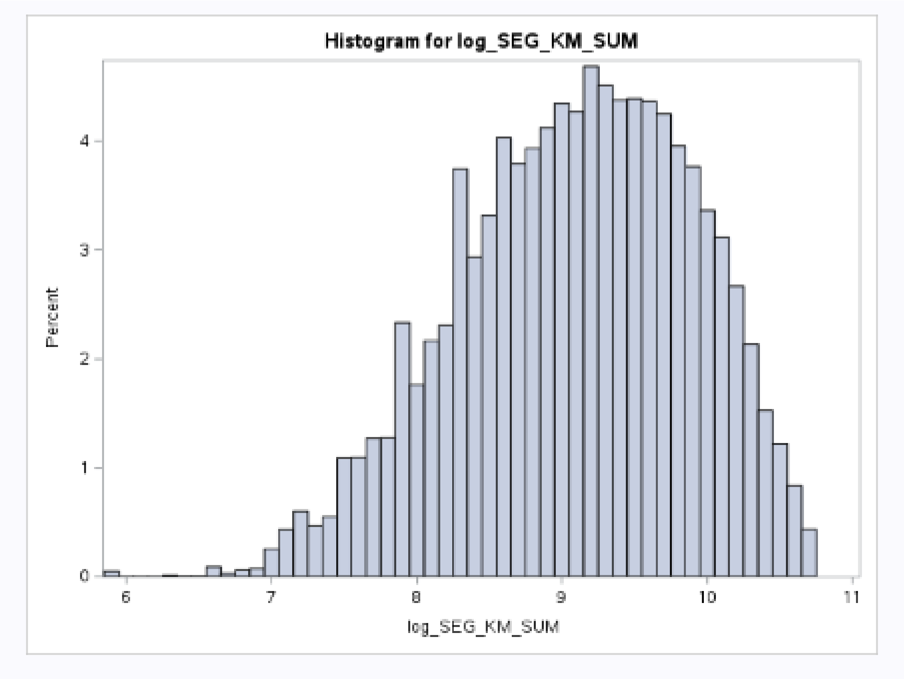


Figure 3.24.17: After using logarithms in SEG_KM_SUM

The histogram of SEG_KM_SUM after applying a logarithmic transformation (Figure 3.24.17) shows a more normalized distribution. The transformation compresses the range of the data and reduces the skewness, bringing the distribution closer to a bell curve. This transformation helps stabilize the variance and makes the data more suitable for regression analysis and other statistical methods that assume normality. The log transformation effectively handles the wide range of values and the presence of extreme outliers, enhancing the interpretability and robustness of subsequent analyses.

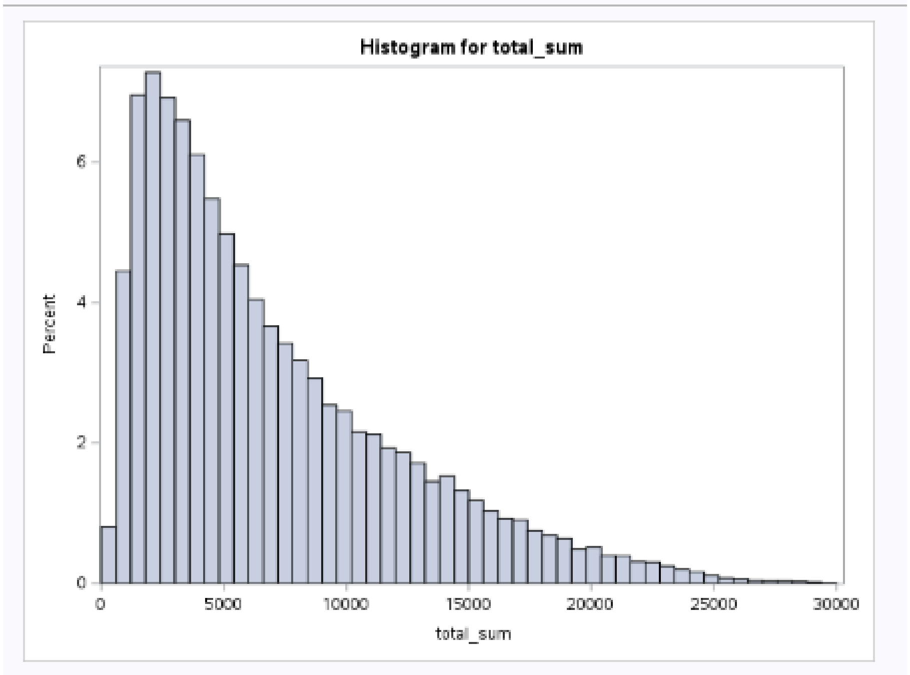


Figure 3.24.18: Before using logarithms in total_sum

The histogram of total_sum before applying a logarithmic transformation (Figure 3.24.18) also shows a right-skewed distribution similar to SEG_KM_SUM. Most customers have relatively low total spending, with a long tail of higher spending values. The right skewness indicates a concentration of lower spending amounts and a gradual decrease in frequency as the total spending increases. This skewness can affect the performance of predictive models and violate assumptions of normality.

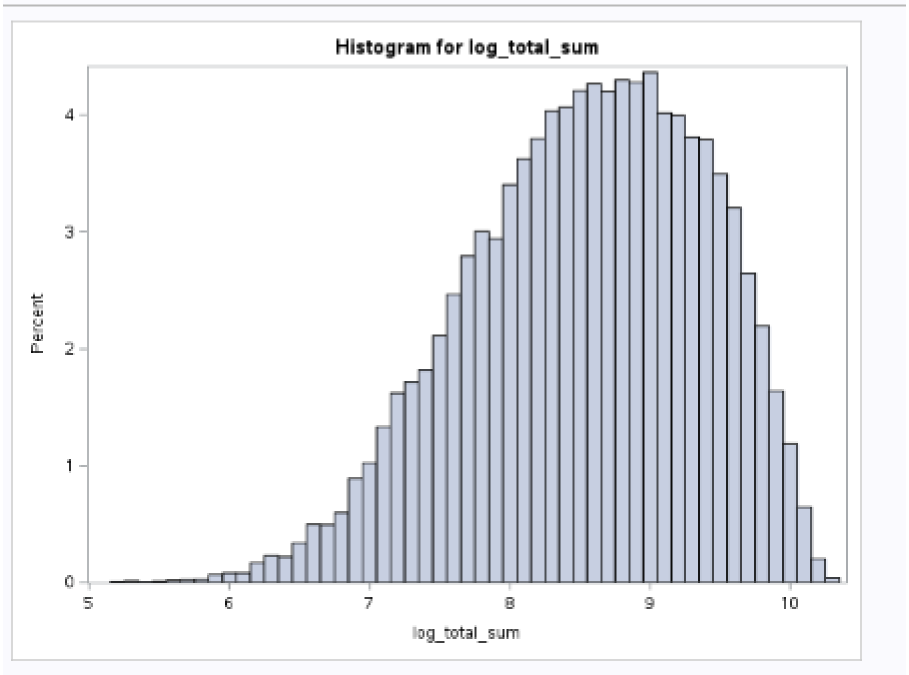


Figure 3.24.19: After using logarithms in total_sum

The histogram of total_sum after applying a logarithmic transformation (Figure 3.24.19) demonstrates a significantly more symmetrical and normalized distribution. The log transformation reduces the impact of extreme values and compresses the data range, resulting in a distribution that is closer to a normal distribution. This normalization improves the suitability of the data for regression analysis and other statistical techniques that require normality. By applying the logarithmic transformation, the data becomes more homogenous, reducing the influence of outliers and improving the robustness of the models.

After Pre-Processing:

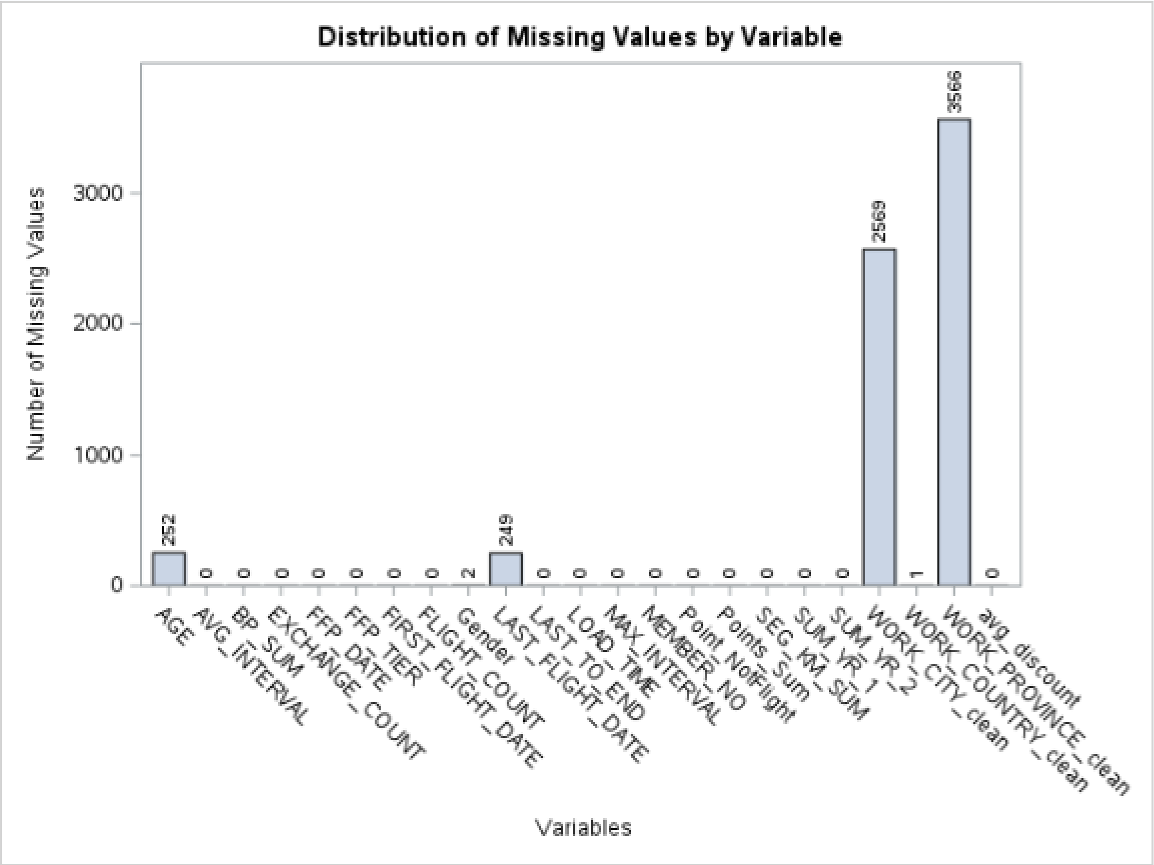


Figure 3.24.20: Updated Distribution of Missing Values by Variable

Post-cleaning, the dataset was reduced to 62,857 entries. This reduction indicates that rows with excessive missing values (more than two) were removed to improve data integrity.

4.0 Conclusion

This assignment underscores the importance of meticulous data pre-processing and the strategic application of data mining techniques in the transportation sector. Through the exploration and preparation of the transportation dataset, we have demonstrated how handling missing values, normalizing data, encoding categorical variables, and selecting relevant features can significantly impact the quality of analytical results.

Our use of a selected data mining technique in SAS Studio has yielded insightful information on the transportation industry, demonstrating the potential to boost service delivery, optimize resource allocation, and increase operational efficiency. The results of the trials and studies show how important it is to have well-prepared data for predictive models to function accurately and dependably.

In conclusion, this assignment has not only highlighted the challenges inherent in data pre-processing but also showcased the transformative power of data mining when applied to real-world transportation data. Moving forward, the lessons learned, and the methodologies developed here can be further refined and expanded, paving the way for more sophisticated and impactful data-driven solutions in the transportation industry.

Individual Assignment

5.0 Azelea Glory Ng Zi-Lin

5.1 Introduction

Aviation plays a huge roll in the international transport sector; its important role is to transport people and goods through over large distances. A lot of information from both internal and external sources is handled in the aviation business. These includes the sales of tickets, flight data, feedback, and social media contacts. The information obtained within this paper helps us to identify important places for improvement. Major areas requiring amendment are productivity, security, and customer satisfaction. Still, they contain enormous amounts of data that handle many other technologies which may confuse issues with data analysis. By using data mining, the process of searching and analysing a large batch of raw data to figure out patterns and useful information to help improve service delivery for the airlines. This paper is about data mining in the aviation industry using the Random Forest algorithm. By using MAE, MSE, and ROC-AUC analysis, we are going to evaluate the challenges faced and the future trends in data mining. The main objective of this paper is to demonstrate how data mining can bring changes in the aviation industry and address the related challenges.

5.2 The motivation for using data mining in the area

Data mining helps to improve scheduling, operations, services, safety, and security in the field of aviation. Useful trends, patterns, and insights to make better decisions are observed from aviation through data mining.

Optimising flight schedules enables the airline to adapt their flight schedule to avoid any disparity with the demand of flights and to be sure that the planes are utilised. This is carried out by analysing past flight schedules, the frequency of bookings, and the traffic flow of tourists during different periods of the year. The airline is able to forecast busy times, popular flight routes, and the precise moment when customers order their tickets by employing data mining. Data mining helps create improved schedules which reduces the number of delays, the cost of fuel, and increase efficiency.

The second factor is **improving customer experience**. Customers expect transport services to fit their expectations and the experience must be optimised. This is because that the airline industry needs to identify and expect the demands of the customers and expectation to provide relevant services. By using data mining to analyse specific data such as experience feedback and their travelling history and preferences enables the airline to offer services like that of loyalty programs, entertainment services, and meals according to the traveller’s tastes.

The last factor is to **improve safety and security**. A huge concern when it comes to the safety of the passengers and the aircraft. The risk assessment and preventive measures against any accidents can be obtained through data analysis. Data mining of maintenance logs, records, flight data and incident reports during flight is subjected to data mining in order to look for safety issues which helps eases to come out with safety and security measures.

5.3 The data mining/machine learning techniques used

The **Random Forest** technique can be highly recommended as one of the most effective and reliable methods of data mining. Recent advancements have made it highly efficient in dealing with lots of information, making it ideal for aviation use. Random Forest is a form of Learning Algorithm which is under Computational Intelligence, and it is formed through Decision Tree or many Decision Trees. In this method, numerous decision trees are developed during the training phase and the mode of the classes (classification) or mean prediction (regression) of the separate trees is provided.

Random Forest can be used for multiple predictive jobs such customer demand, prediction of delayed flights, and maintenance. The algorithm being able to handle large input of variables makes it suitable for complicated and data from the aviation industry.

Random Forest is especially helpful for aviation data mining because it has many benefits. First of all, it **eliminates overfitting and improves forecast reliability** by averaging the output of several trees, giving it high accuracy. By efficiently managing noisy data and missing values to ensure consistent performance on different datasets, this shows that the algorithm is robust. Furthermore, because Random Forest yields feature importance scores, which reveal which variables have the greatest influence on the results, it is very interpretable. Random Forest is a useful tool for collecting important information from aviation data because of its great interpretability, robustness, and accuracy.

Another important application of Random Forest is **flight delay prediction** because it predicts the delays based on the weather condition, traffic, and prior delay data. With this approach, airlines get the chance to inform passengers within less time and also manage their schedules in a proper way. Random Forest can be used to analyse the information related to the engine and aircraft such as performance data, flying hours, and the logs of the number of flights and the maintenance that has been done on the aircraft. This means it is possible to predict when future maintenance may be needed by the equipment pieces. This helps in minimising regular repair incidences hence minimising their downtime due to these repairs.

To make a final prediction, the Random Forest algorithm creates several decision trees and combines their outputs. Bootstrap sampling is the first step, in which replacement is used to construct random subsets of the training data. This means that some information may appear more than once, while others may not appear at all. A decision tree is created for every subset using a random feature selection, creating variation between the trees. Each tree gives a vote for a class in classification tasks, and the class with the highest number of votes is selected. The final output for regression tasks is the mean predicted across all trees.

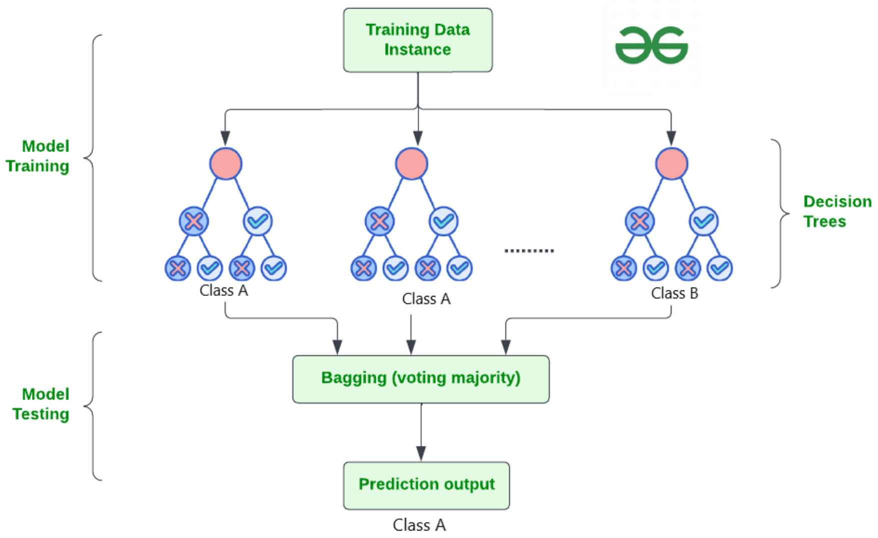


Figure 5.1.1: Illustration of Random Forest algorithm building multiple decision trees and combining their outputs (GeeksforGeeks, 2024)

Two key concepts from ensemble learning are used by Random Forest: feature randomness and bagging (also known as bootstrap aggregating). Using random sampling with replacement, bagging involves dividing the original dataset into several subsets, each of which is then used to create a different decision tree. By choosing a random subset of features at each split in a tree, feature randomness increases diversity even more. By reducing the correlation between trees and ensuring the distinct qualities of each tree, this method creates a more robust model as a whole.

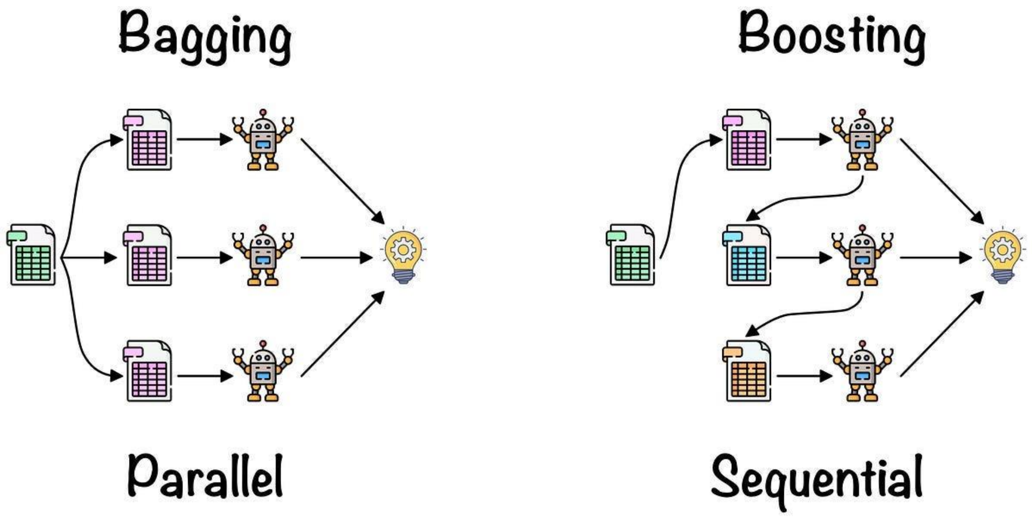


Figure 5.1.2: Bagging process in Random Forest (Analytics Vidhya, 2024)

Random forest is a very powerful technique when it comes to aviation data mining as it is able to deal with large amounts of data and is highly accurate and stable. It is more efficient for the airlines to use an accurate model of Random Forest in decision making process on passenger demand, scheduling of flights and maintenance demands. Even though it is very beneficial to use, it is important to guard privacy and data quality issues during the implementation process. Random Forest in the topic of aviation has increasing potential as more advances are made in the field of machine learning.

5.4 The evaluation method of the data mining model

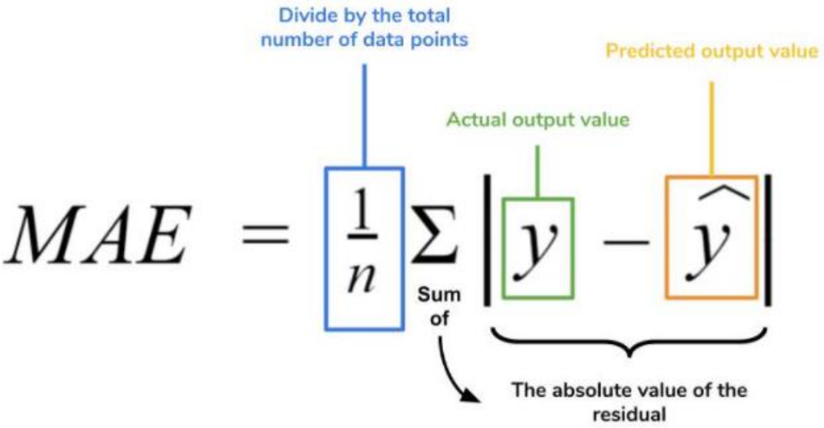

Divide by the total number of data points
Predicted output value
Actual output value
MAE = 1/n Σ |y − ŷ|
Sum of
The absolute value of the residual

Figure 5.2.1: Formula for MAE

Mean Absolute Error (MAE):

Without considering the tendency of the errors, MAE calculates the average magnitude of the errors in a series of predictions. It is the mean, with each individual difference carrying equal weight, of the absolute differences between the observed and predicted values over the test sample. MAE is a simple metric that can be used for activities such as estimating passenger demand since it gives a quickly interpreted measure of prediction accuracy.

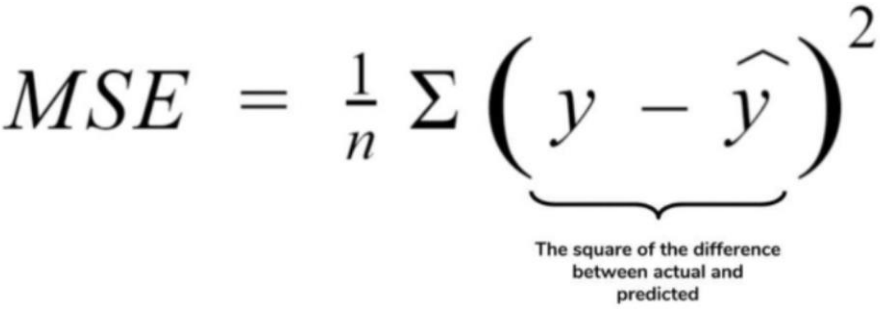


Figure 5.2.2: Formula for MSE

The difference between the expected and actual numbers is measured by the Mean Squared Error (MSE), which is the average of the squares of the errors. In comparison with MAE, MSE is more prone to outliers because it assigns a higher weight to larger errors. It is helpful to assess the precision of ongoing predictions, such as those for aircraft fuel consumption or maintenance requirements. A model that fits the data better is shown by a lower MSE.

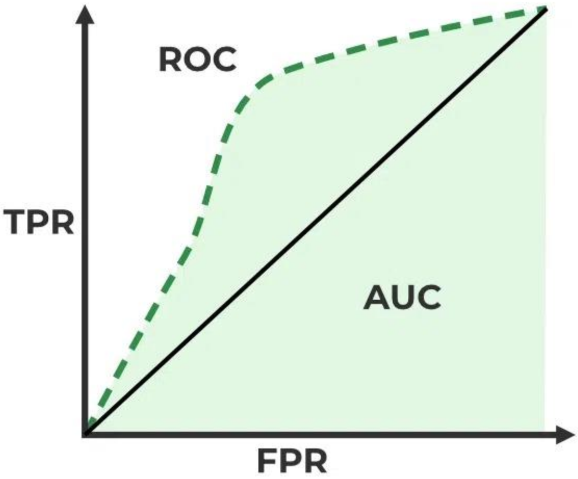


Figure 5.2.3: ROC-AUC Classification Evaluation Metric (GeeksforGeeks, 2024)

When a binary classification system's discrimination threshold is changed, its diagnostic ability is shown graphically using the Receiver Operating Characteristic (ROC) curve. The degree or measure of separability is represented by the Area Under the Curve (AUC), which shows how well the model can differentiate between groups. For binary classification problems in aviation, such as predicting flight delays (delay/no delay), the ROC curve and AUC are helpful. A higher AUC value shows that the model does a better job at differentiating between the two classes. False positive rate (FPR) is the opposite of true positive rate (TPR).

An example application is airlines may transform models to give better predictions by using an MAE that can help to measure the average error magnitude regarding the forecasts made on passenger demand. Having assessed these elements and regarding estimated and actual maintenance requirements, the MSE can detect any considerable disparities that need an intervention in order to adjust the maintenance schedule. Real life case: The performance of the model can be measured by Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) for classification of the planes as delayed or not to improve the operational efficiency of the airlines for the enhanced communication with the passengers.

5.5 New/prospective opportunities produced by the analysis

To meet the individual's needs, **personalised passenger experience** such as tailoring services and communication. Using the data from booking patterns, in-flight behaviour, and feedback the airline can come up with a more engaging and a comfortable journey for its passengers. Frequent flyers would be offered loyalty points and customers would be offered promotions for many destinations.

A priority in the aviation industry is that of **safety and security**. By using data mining, it helps the airlines to analyse huge amounts of data, which enables to look for risks and provide ways to prevent it. Using sensor data from the aircraft, maintenance needs, and incident reports we can monitor what is needed to be done and how they can inhibit any safety issues from happening.

Growing relevance of **sustainability** in the aviation sector. Data mining may optimise fuel utilisation and emission reduction to be significant on sustainability drives. By analysing flight data, airlines could understand inefficiencies of route selection as well as flight schedules which would help them adopt more fuel-efficient practices while also knowing passengers' taste on environmentally safe products will lead them into making decisions that are aimed at sustainable travels including a program about carbon offsets.

These opportunities underscore that the potential outcomes when we use data to dig out information in aviation are life-changing because they enhance passenger satisfaction alongside making operations more efficient which leads to safety advancement and environmental conservation efforts, using big data companies in danger of failing will be avoided thanks to flight statistics analysis.

5.6 Challenges faced in this area

Apart from providing the aviation sector with several advantages, data mining faces some major obstacles that need to be tackled if its application would be successful. This indicates that **data integration** is one major concern in this case. Most of the time, the aviation data is dispersedly kept in various systems and structures, thus it becomes hard to integrate. Different departments within an airline; for example, booking, operation and maintenance might use different systems which may not easily readjust themselves as compatible units.

Another challenge is **model interpretability** when it comes to data science. Models developed using sophisticated approaches to mining complex data, such as Random Forest algorithm, can end up being too difficult to comprehend. Also, stakeholders could be having a difficult understanding behind different decisions that are made. Consequently, "black box" perception associated with certain models can act as an impediment towards their acceptance and integration into industry since decision makers would be unwilling to heed insights which they do not fully understand.

Real-time data processing is a huge challenge. It is required in the aviation industry that real-time data processing is used to carry out decisions. The processing of large amounts of data in real-time is demanding and is intensive. Delays could disrupt flight schedules or informing passenger on delays. Being able to analyse the weather conditions, air traffic, and passenger data could aid in flight management. However, implementing this system would prove to be complicated and costly.

The final drawback is that the **ever-changing face of aviation operations** tends to be highly challenging since the sector is **very volatile**; it changes rapidly as a result of various conditions such as weather, air traffic control issues as well operational disruptions. Hence, in order for data mining models to be always relevant they have to keep being updated each time there is change. Keeping data mining models up to date with the latest data is difficult because it involves continuous monitoring and retraining which can consume a lot resources. If unexpected events such as health crises or geopolitical concerns cause sudden changes in travellers' movements, the accuracy of a model that forecasts demands using data from the past may be inaccurate.

5.7 Future direction(s) of research in this area

Data mining in the aviation industry has a lot of possibilities in future development. Combining **real-time data analytics with the Internet of Things (IoT)** is one potential path. IoT devices are able to gather real-time data from a range of sources —including airport systems, passenger devices, and aircraft sensors— to offer current information for dynamic decision-making. Improving operational effectiveness, boosting safety, and customising the traveller experience are all possible with this integration. For example, by giving real-time updates, IoT sensors can be used by smart airports to track passenger movement, allocate resources properly, and improve overall operations.

To deal with the rising concerns about data security, **better data privacy** approaches must be developed. Federated learning and differential privacy techniques may be studied more in the future, which will allow data analysis while maintaining personal privacy. Airlines can gain valuable insights from these tactics without risking customer confidence or complying with rules. Federated learning, for example, can be used to create machine learning models across decentralised data sources while maintaining local data, ensuring security and privacy.

5.8 Conclusion

In summary, through this paper, it is very clear that data mining can significantly transform the aviation industry to ensure better operational capabilities, customer satisfaction, as well as safety. Such algorithms as Random Forest are able to address important tasks like passenger demand predicting, flight delay time identifying, as well as possibility of further maintenance calculating. However, with all its advantages, there remain the problems of data integration, model interpretability, data processing in real-time, as well as the fluctuating nature of aviation processes. Some of the areas for further research include IoT real time incorporation and incorporation of real time analysis, enhancing machine learning approaches for prediction and other methods that can ensure the improvement of privacy. When these challenges are addressed and the future directions considered, the aviation industry can benefit from data mining to foster increased innovation and operational efficiency.

5.9 References

Author links open overlay panelTri Noviantoro, environment, A. the current competitive, Bellizzi, M. G., Chang, Y. H., Chen, C. F., Chou, C. C., Dolnicar, S., Gudmundsson, S. V., Guo, Y., Hu, K. C., Hussain, R., Jiang, H., Jou, R. C., Kumar, S., Kuo, M. S., Laming, C., Leon, S., Lim, S. S., Liou, J. J. H., … Wu, H. C. (2021, October 14). *Investigating airline passenger satisfaction: Data Mining Method*. Research in Transportation Business & Management. https://www.sciencedirect.com/science/article/abs/pii/S2210539521001097

Avcontentteam. (2024, March 18). *Tree based algorithms: A complete tutorial from scratch (in R & python)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/

Bhandari, A. (2024, April 23). *Guide to AUC ROC curve in machine learning : What is specificity?*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Bhatti, S. (2020, October 3). *The art and science of data mining in the aviation industry*. LinkedIn. https://www.linkedin.com/pulse/art-science-data-mining-aviation-industry-sarosh-bhatti

Britton, T. (2023, October 20). *What is importance of data mining in aviation SMS implementations*. What Is Importance of Data Mining in Aviation SMS Implementations. https://aviationsafetyblog.asms-pro.com/blog/the-importance-of-data-mining-in-sms-programs#:~:text=What%20Is%20Data%20Mining%20in,the%20organization%27s%20risk%20management%20processes.

DeLaura, R., & Reynolds, T. (n.d.). Big aviation data mining for robust, ultra-efficient air transportation ... https://nari.arc.nasa.gov/sites/default/files/attachments/Reynolds_Abstract.pdf

Donges, N. (n.d.). *Random Forest: A complete guide for machine learning*. Built In. https://builtin.com/data-science/random-forest-algorithm

Gavrilovski, A., Jimenez, H., Mavris, D. N., & Rao, A. H. (n.d.). (PDF) challenges and opportunities in flight data mining: A review of the state of the art. https://www.researchgate.net/publication/290194327_Challenges_and_Opportunities_in_Flight_Data_Mining_A_Review_of_the_State_of_the_Art

Glaze, R. (2023, July 3). *Analyzing airline data: A Data Mining Project*. LinkedIn. https://www.linkedin.com/pulse/analyzing-airline-data-reid-glaze

Helsen, S. (2023, April 19). *The 3 big challenges with big data in aviation*. Hexagon Safety, Infrastructure & Geospatial blog. https://sigblog.hexagon.com/the-3-big-challenges-with-big-data-in-aviation/

Howell, C. (2023, December 19). *How to do predictive risk management data mining in Aviation SMS*. How to Do Predictive Risk Management Data Mining in Aviation SMS. https://aviationsafetyblog.asms-pro.com/blog/best-data-mining-methods-predictive-aviation-safety-risk-management

Küsbeci, P., & Burak, M. F. (n.d.). The Impact of Data Mining in The Aviation Industry. https://acikerisim.kapadokya.edu.tr/xmlui/handle/20.500.12695/196

M, P. (2024, May 27). *A comprehensive introduction to evaluating Regression Models*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/#:~:text=Mean%20Absolute%20Error%20(MAE)%20is,and%20is%20easy%20to%20interpret.

Machine learning random forest algorithm - javatpoint. www.javatpoint.com. (n.d.). https://www.javatpoint.com/machine-learning-random-forest-algorithm

Narkhede, S. (2021, June 15). *Understanding AUC - roc curve*. Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

NASA. (n.d.). *NASA data mining algorithms*. NASA Data Mining Algorithms | SKYbrary Aviation Safety. https://skybrary.aero/articles/nasa-data-mining-algorithms

Pagels, D. A. (2015, March). Aviation Data Mining - University of Minnesota, Morris Digital well. https://digitalcommons.morris.umn.edu/cgi/viewcontent.cgi?article=1023&context=horizons

Polanitzer, R. (2022, March 23). *The minimum mean absolute error (MAE) challenge*. Medium. https://medium.com/@polanitzer/the-minimum-mean-absolute-error-mae-challenge-928dc081f031

R, S. E. (2024, May 23). *Understand random forest algorithms with examples (updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Random Forest algorithm in machine learning. GeeksforGeeks. (2024, February 22). https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

Ravin. (2024, January 25). *Auc Roc Curve in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/auc-roc-curve/

Simplilearn. (2023, November 7). *Random Forest algorithm*. Simplilearn.com. https://www.simplilearn.com/tutorials/machine-learning-tutorial/random-forest-algorithm

Sj&ouml;blom, O. (1970a, January 1). *Data mining challenges in the management of Aviation Safety*. SpringerLink. https://link.springer.com/chapter/10.1007/978-3-662-45526-5_21

Sj&ouml;blom, O. (1970b, January 1). *Data mining in promoting aviation safety management*. SpringerLink. https://link.springer.com/chapter/10.1007/978-3-319-10211-5_19

What is Random Forest?. IBM. (2021, October 20). https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,Decision%20trees

6.0 Churilov Mikhail

6.1 Introduction

Data Mining is an important area of data science that focuses on extracting knowledge from large amounts of data. In today's world, with the rapid growth of data volumes, mining methods have become key tools for business, science and various industries. They allow us to identify hidden patterns, trends and anomalies that cannot be detected by traditional methods of analysis.

With the development of technology and the increase in data volumes, it has become necessary to use more complex and powerful analysis methods. Traditional data processing methods, such as simple statistics and visualization, are no longer able to cope with the volume and complexity of modern data. Data mining provides tools for automatically searching for significant structures and patterns in data, which allows you to make informed decisions and predict future events.

In this essay, the regression method, one of the main and most frequently used methods of data mining, will be considered in detail. Regression is used to model and analyse dependencies between variables, as well as to predict quantitative values based on historical data. The introduction to regression includes a description of the main types of regression, such as linear, polynomial, and logistic regression, as well as vector-supported regression.

Regression analysis has found wide application in various fields such as economics, medicine, engineering and marketing. In particular, in the field of air transportation, regression models are used to predict passenger traffic, optimize routes and improve the quality of customer service. This essay will consider a specific case of using regression analysis in aviation, including all stages of working with data, from preprocessing to evaluating the results of the model.

6.2 The motivation for using data mining in the area

Data mining meets the needs of:

Extracting hidden data: Large data sets often contain important information that may be overlooked in traditional analysis. Intelligent analysis helps to identify these hidden patterns and use them to make informed decisions.

Forecasting: In accounting, finance and healthcare systems, data mining techniques are used to predict future events, which allows you to anticipate potential problems and take action in advance.

Improving efficiency: The use of deep data analysis methods can significantly increase the productivity of companies, improve the quality of solutions and reduce costs by detecting weaknesses and optimizing processes.

Personalization: In marketing and customer service, data mining helps segment the audience and offer personalized offers, which increases customer loyalty and improves their user experience.

Anomaly detection: Mining techniques are used to identify anomalies in data, which is especially important in banking and cybersecurity, allowing you to reduce financial losses and prevent data leaks.

Decision support: Analytical departments can use the results of data mining to provide marketing and other departments with more accurate and detailed information, which contributes to making more informed decisions.

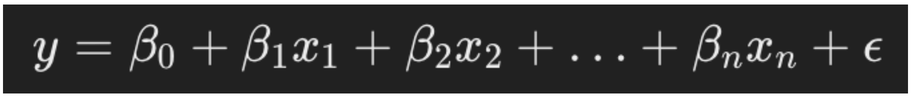
Data mining methods are used in a wide variety of fields, including business, healthcare, science, education and public administration. In this essay, we will focus on one of the key methods — regression analysis, its principles, data preprocessing methods, limitations and advantages, as well as a specific example of its application in the field of air transportation.

6.3 The data mining/machine learning techniques used

Regression analysis is one of the fundamental methods in the field of data mining. Its main purpose is to model and analyze the dependence of one variable on others. Using regression analysis, it is possible not only to identify existing dependencies, but also to predict future values based on historical data. In this section, we will take a detailed look at the main types of regression, their methods of operation, as well as the cases in which they are most effective.

There are four main types of regression. The first one is linear regression.

Method description: Linear regression is the simplest and most widely used regression analysis method. It models the relationship between the dependent variable y and one or more independent variables x using a linear equation. The linear regression equation can be written as follows:



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Figure 6.1.1

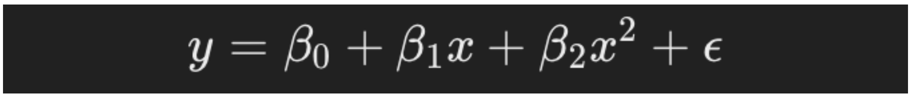
where β0 is a free term, β1, β2, βn are regression coefficients, x1, x2, xn are independent variables, and ϵ is the error of the model.

The method works: Linear regression uses the least squares method to minimize the sum of the squares of the differences between the actual values of the dependent variable and the values predicted by the model. This allows us to find the optimal values of the coefficients β.

Application: Linear regression is widely used in cases where there is a linear relationship between variables. It is used in economics to analyze supply and demand, in medicine to assess the impact of various factors on health, in marketing to predict sales and analyze consumer behavior.

The second type of regression is polynomial regression.

Method description: Polynomial regression extends linear regression by adding polynomial terms to the model. This allows you to model nonlinear dependencies between variables. The second-order polynomial regression equation can be written as follows:

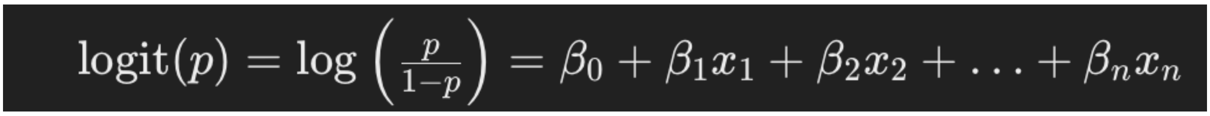


$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Figure 6.1.2

Method operation: Similar to linear regression, polynomial regression uses the least squares method to find coefficients that minimize model error. However, unlike linear regression, it allows for more complex dependencies by including polynomial terms.
Application: Polynomial regression is used in cases where the relationship between variables is nonlinear. For example, it is used in physics to model dependency curves, in biology to analyze population growth, and in economics to model complex financial dependencies.

Method description: Logistic regression is used for classification problems where the dependent variable takes binary values (for example, 0 or 1). Unlike linear regression, logistic regression models the probability of an event occurring. The logistic regression equation can be written as follows:



$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Figure 6.1.3

where p is the probability that the dependent variable is 1.
Method operation: Logistic regression uses the maximum likelihood method to estimate the coefficients of the model. It converts the output values of the logistic function into the range [0, 1], which allows them to be interpreted as probabilities.
Application: Logistic regression is widely used in medicine to predict the probability of disease, in marketing to analyze customer behavior (for example, the probability of buying a product), in finance to assess the probability of default on loans.

Method Description: Support Vector Regression (SVR) is an extension of the Support vector Machine (SVM) for regression problems. SVR uses a hyperplane to predict the values of a dependent variable, aiming to minimize the prediction error.
The method works: SVR finds a hyperplane that minimizes the prediction error, while allowing some error within a certain threshold (epsilon). This allows the model to be resistant to outliers and noise in the data.

Application: SVR is used in cases where high accuracy of forecasts is required and the data has complex structures. For example, it is used in financial markets to predict stock prices, in energy to predict energy consumption, and in bioinformatics to analyze genetic data.

Before applying regression analysis, it is important to conduct thorough data preprocessing to ensure the quality and reliability of the models. This process includes several key steps:

1. Data Cleanup: Deleting or correcting missing values, duplicates, and anomalies. For example, if data on passenger traffic is missing for certain days, you can use imputation methods to restore the missing values.
2. Normalization and standardization: Converting data to a single scale. This is especially important for data scale-sensitive methods such as linear regression. For example, normalization of data on ticket prices and the number of passengers avoids the dominance of one of the features in the model.
3. Categorical Variable Conversion: Converting categorical data into numeric form using methods such as one-hot encoding. For example, converting months and days of the week into numerical signs to account for seasonal and daily fluctuations in passenger traffic.
4. Feature selection and creation: Identify key variables to be used in the model and create new features based on existing data. For example, adding features such as weather conditions, holidays and events that may affect the number of passengers.

6.4 The evaluation method of the data mining model

Now let's look at the application of various types of regression analysis using specific examples to better understand in which cases each of them is most effective.

Linear regression is especially useful when there is a direct proportional relationship between variables. For example, in economics, linear regression is used to analyze the dependence of income on education and work experience. In medicine, it can be used to assess the effect of age and body mass index on blood cholesterol levels.

Polynomial regression is used in cases where the data shows a non-linear relationship. For example, in ecology, to model the dependence of plant growth on time, where growth does not occur linearly, but has certain stages. In engineering, polynomial regression can be used to analyze complex dependencies between parameters and characteristics of materials.

Logistic regression is widely used in binary classification problems. In medicine, it is used to predict the likelihood of diseases based on various risk factors. In marketing, logistic regression helps to estimate the probability that a customer will make a purchase based on their behavior and demographic data.

SVR is used in cases where the data has complex structures and requires high accuracy of predictions. For example, in financial analysis, SVR is used to predict stock prices, where data can be very noisy and have complex dependencies. In energy, SVR can be used to predict energy consumption based on historical data and weather conditions.

Regression analysis is a powerful tool for modeling and analyzing dependencies between variables. Various types of regression offer solutions for a wide range of problems, from simple linear dependencies to complex nonlinear and classification problems. Understanding the principles of operation and areas of application of each method allows data analysts to choose the most appropriate tools for specific tasks, providing accurate and reliable forecasts. In the next section, we will look at the advantages and limitations of regression analysis, as well as a specific case of its application in the field of air transportation.

6.5 New/prospective opportunities produced by the analysis

Optimized Route Planning in aviation mainly is to enhance the passenger experience while improving the operational efficiency. By leveraging insights from passenger demand pattern and preferences (Kousik et al., 2020), aviation can arrange route planning strategies to meet up the passenger or customer needs. This can lead to the introduction of new routes or increased frequency on existing popular routes and improve customer satisfaction and loyalty.

6.6 Challenges faced in this area

Complexity of Factors: By analyse and collect data about travel behaviour analysis, analysist will find out there are multiple part that each of the customer and traveller have different behaviour. For example, the parts that will influence of the customer's journey are the ticket pricing, flight schedules, airline company, airport facilities, sitting preference, travel purpose etc. Based on these factors, analysist needs to be very professional and needs to take more time to research about the factors of travel behaviour analysis.

6.7 Future direction(s) of research in this area

Utilising more advanced machine learning algorithms for statistical prediction. Some of the technologies such as the deep learning technologies, neural networks, amongst others, can assist in estimating the subsequent travel patterns, demand of passengers and even potential hitches. These models are capable of identifying intricate patterns in the large volume of complex data and yield accurate and elaborate projections. Maintenance planning, flight scheduling, and resource allocation can be improved significantly. Deep learning models, for example, can forecast passenger demand by analysing previous booking data, economic indicators, and social media trends. This enables airlines to improve customer satisfaction and optimise their services.

6.8 Conclusion

Regression analysis is a powerful and flexible tool for analyzing and modeling dependencies between variables. Various types of regression, such as linear, polynomial, logistic and vector-supported regression, offer solutions for a wide range of tasks. Understanding the principles of operation and areas of application of each method allows data analysts to choose the most appropriate tools for specific tasks, providing accurate and reliable forecasts. The application of regression analysis in the airline industry illustrates how these methods can be used to predict and optimize business processes, improving decision-making and increasing the efficiency of operations.

6.9 References

Aggarwal, C. C., & Reddy, C. K. (2013). Data Clustering: Algorithms and Applications. CRC Press.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Xu, R., & Wunsch, D. (2009). Clustering. Wiley-IEEE Press.

Tan, P. N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.

Regression Methods:

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). Wiley.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). Wiley.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models (5th ed.). McGraw-Hill Irwin.

7.0 Jane Lee

7.1 Introduction

Transportation is the movement of people and things between locations as well as the several ways that this movement is carried out (The Editors of Encyclopaedia Britannica, 2024). Transportation has included buses, train services, aviation, personal automobiles, pedestrian, and boats transportation (*Types of Transportation – RHIHub Toolkit*, n.d.). Perhaps some people will think that transport this topic is unimportant and that there is nothing to worry about it, but community mobilization will be hampered if people have lack access in reliable transportation (Firdausi et al., 2023).

Based on this big topic, the research area that I wanted to explore more is about travel behaviour analysis in aviation. Most of the people would like to travel around by using transportation.

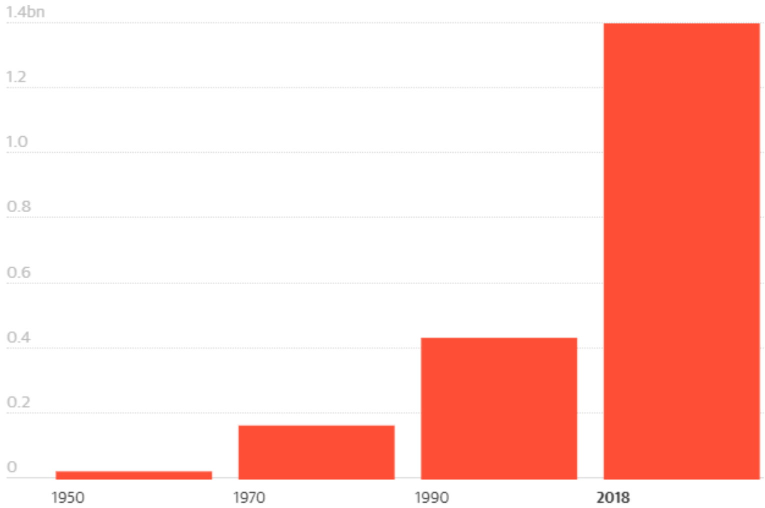


Figure 3.1.1: Global tourism research since the end of second world war (Blackall, 2019b)

Based on the Figure 3.1.1 can be seen that the highest record of travelling during school holiday could be 1.4 billion in the research year between 1950 to 2018. From this statement, can know that most of the people likes to travel around.

Research in travel behaviour analysis is to learn and study about the sociology of people making decisions about travelling for example the mode of transportation, the frequency of travel, the places to travel and the travel time. The reason to research about the travel behaviour of people or individuals is to have better understanding about the factor of decision in travelling.

By research in this travel behaviour analysis, will provide more understanding about people or individuals before, during and after travelling.

The outcome of analyse, researchers could provide the data that had been analysed to the industry that related to travel, industry could more understand travellers’ perspective and improve the transportation system and related services by having the data to do decision making.

7.2 The motivation for using data mining in the area

Data mining is an examining vast volume of data and datasets in order to extract useful intelligence that can assist organisations in resolving issues, identifying patterns, reducing risks, and seizing new possibilities (Simplilearn, 2023).

Data mining had been chosen in analyse travel behaviour analysis in aviation industry is due to the fact that in aviation industry not only just have people in surrounding, but it also contains a bunch of data in their system or dataset. Based on all of the data that had been collected in the dataset or system, it usually will be complicated and large, so if aviation industry wants to have improvement, aviation industry or related departments needs to analyse data to know more information based on every traveller.

Firstly, the motivations are optimizing flight schedules and routes (Chang & Schonfeld, 2004). By using data mining, airline can optimize flight schedules and routes by analysing travel patterns and passenger demand. Besides that, aviation industry could be based on data mining to know more about the peak travel hours (usually is school holiday session), locations that traveller usually chose and seasonal fluctuations and modify the schedule or prices, so aviation industry could gain demand from this method.

Secondly, the other motivation of use data mining in this area is enhance customer or traveller experience and satisfaction. Airlines are able to enhance the whole travel experience by tailoring services based on feedback, travel history, and customer preferences. This might include tailored promotions, focused advertising efforts, and in-flight amenities.

Thirdly, enhance the safety and security for customers is important in the motivation of use data mining. Safety and security is a way of motivation because customers and travellers will still concern about the risk of the safety and security when travelling by airplane. By examining the data, data mining could spot some security risks and unusual patterns. Therefore, data mining methods could let aviation industry know more about the security threads and could promptly make changes to let customer and travellers feel at ease.

7.3 The data mining/machine learning techniques used

Data mining techniques are employed by airlines to comprehend client preferences, forecast demand, pinpoint areas for cost reduction, and improve safety and security measures (Pankhania, 2023).

For instance, analysist may divide up travellers into several market segments by using data mining to find groups of travellers that share traits, habits, or interests. Data mining may also be used to find abnormalities, outliers, or shifts in travel patterns that could point to issues or opportunities (*How Can You Use Big Data to Improve Travel Behavior Analysis?*, 2023).

Neural Networks (Convolutional Neural Networks - CNNs)

A Convolutional Neural Network (CNN) is a type of deep learning algorithm especially for image processing and recognition applications. CNNs require lesser preparation than other classification models because they can automatically learn hierarchical feature representations from raw input pictures (*Convolutional Neural Network: Benefits, Types, and Applications*, 2023).

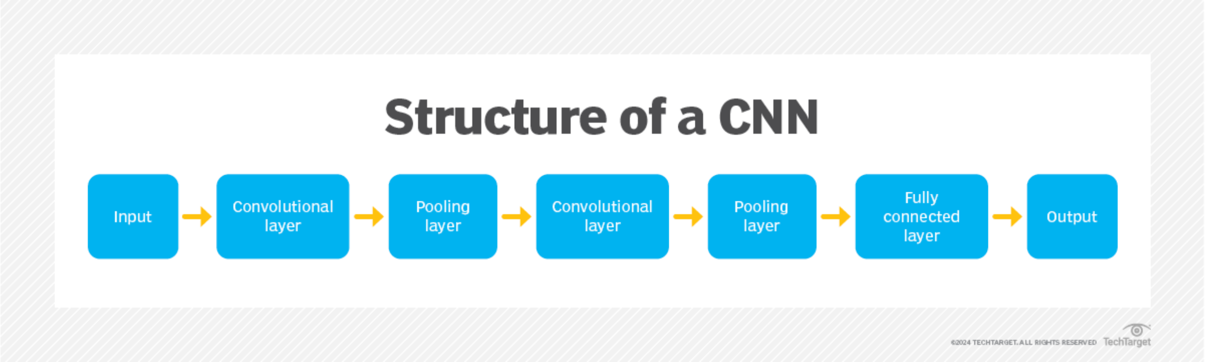


Figure 3.3.3: Structure of CNN (Craig & Awati, 2024)

Based on the Figure 3.3.3 can clearly know that CNN has layers and people will be divided it into three categories, which was convolutional layers, pooling layers and fully connected layers. The CNN's complexity rises as data moves through these layers, enabling it to detect ever more abstract characteristics and greater areas of a picture.

Convolutional layer: Is most of the calculation that will take place in CNN and the basic building element. Convolutional layer will explore the receptive field of an input picture using a filter or kernel, which is a tiny matrix of weights, to look for the characteristics. For instance, in this area airline could analyse CCTV footage to understand passenger flow patterns, identify the congestion points and optimize terminal layout (Mishra, 2021).

Pooling layer: Is the crucial part of CNN, which are used in deep learning. The task in this later is to minimise the supplied data's spatial dimensions while preserving its most crucial elements (Dremio, 2023). For instance, analysist could summarize the data from passenger movement analyses to identify peak travel times and streamline operation.

Fully connected layer: This layer is to classifies the input into the appropriate label by connecting the data derived from the preceding phases (including convolutional and pooling layer) to the output layer (Deepanshi, 2023). For instance, based on this layer analysist could predict passenger behaviour based on the past data including destinations or preferences for certain services.

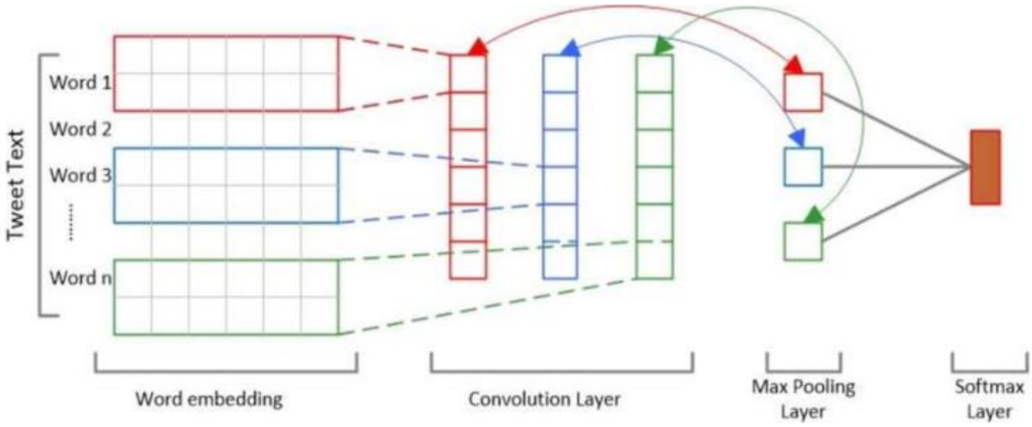


Figure 3.3.4: General architecture of CNN model (Tayaba et al., 2023)

Based on the Figure 3.3.4 that is from a research article about Transform Customer Experience from Twitter message in the Airline Industry. Basically, including words and images could be analyse using CNN, and all of the process will go through in convolution layer, pooling layer and fully connected layer.

7.4 The evaluation method of the data mining model

Cross Validation (k-fold)

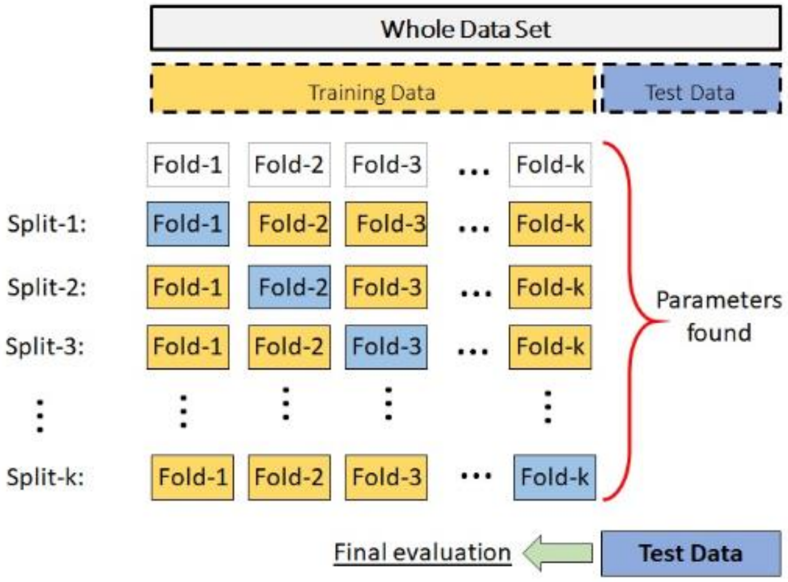


Figure 3.4.3: The concept of Cross Validation (Sevinç, 2022)

In machine learning, cross validation is a technique used to assess a model's performance on omitted or unseen data. In Cross Validation will have various types of cross validation techniques, there are including Holdout Validation, LOOCV (Leave one out cross validation), Stratified Cross Validation and K-Fold Cross Validation. Among these, K-Fold Cross Validation will be more suitable for aviation (Melanie, 2024).

In the concept of K-Fold Cross Validation, it involves dividing the available data into many subsets or folds.

The process of K-Fold (Brownlee, 2023):

1. The dataset will be randomly shuffled and split the data into K folds.
2. For each fold iteration:
    - One of the folds will be used as the validation set
    -The remaining folds, K-1 will be used to train the model
3. The model's performance is evaluated on the validation set.
4. This procedure is iterated several times, with a distinct fold serving as the validation set each time (GeeksforGeeks, 2023a).
5. The results from each fold are averaged to provide a final performance estimate.

K-Fold Cross Validation is critical in aviation for some reasons.
1. Robustness: It make sure that the model’s performance is independent on a single train-test split. This is crucial for application where data variability is high including flight delay prediction and customer segmentation.
2. Generalization: It helps determine how well the model performs on new, unseen data, which is essential for making reliable in dynamic aviation environments.

Therefore, in the aviation industry, Cross Validation, especially in K-Fold Cross Validation is a powerful method in evaluating the data mining models. It could ensure the model robustness and generalizes well to the new and unseen data. This technique is also useful for tasks including customer segmentation, flight delay prediction, route preference analysis and customer satisfaction prediction, for enhancing the operational strategies and improve the passenger experience and satisfaction in aviation.

7.5 New/prospective opportunities by the analysis

Dynamic Pricing Strategy is a data-driven approach to adjust or modify the ticket prices in real-time in order to maximise the income and match the demand (FareIntelligence, 2023b). By predicting demand fluctuations based on historical patterns and external factors, such as holidays, airlines can adjust ticket prices dynamically to maximize revenue while ensuring competitive fares for passengers.

Before adjusting the ticket prices, airlines first need to analyse and collect historical and real-time data to forecast variation in demand depending on variables such as holidays. Predictive modelling and scenario analysis help anticipate future demand, while elasticity assessment and dynamic adjustments optimize pricing. Customer segmentation tailors' prices for different groups, and competitive analysis ensures fares remain attractive. By using these strategies, it could enhance the revenue and customer satisfaction by continuously observe the market conditions and adjusting the price at the same time.

Example of the timeline for adjusting ticket prices (Minhas, 2023):

1. **Holiday Seasons:** Airlines could predict higher demand during holidays and adjust prices according to the maximize revenue while still filling the seats.
2. **Last-Minute Bookings:** By analysing the booking patterns, airline can offer last-minute deals to fill the remaining seats, balancing between the maximizing revenue and minimizing the empty seats.
3. **Event-Based Travel:** Usually if there are a big festival or event been held in some county (for example: Olympics), airlines can predict the demand and dynamically price tickets to capture the maximum revenue.

In summary, dynamic pricing strategies leverage advanced data analytics to optimize ticket pricing in the real-time, aim in balance the revenue goals with customer satisfaction. By using this strategy, aviation could adapt the changes very fast when the market is changing, and aviation could also maximum the financial performance and at the same time can improve the passenger experience.

7.6 Challenges faced in this area

Data mining in travel behaviour analysis although gain a lot of benefits to the aviation industry, but it has challenges while analysing. The challenges including:

Data Availability and Quality:

-By collecting data in aviation industry, the data that collected will be limited and fragmented, due to the fact that all of these data needs to relies on airlines or air traffic management system. Besides, sharing the data to analysis it could be hard due to privacy concern and usually customer and airline company will have data sharing agreements, therefore analysis can't get appropriate and complete data easily.

-According to (Wolf, 2022), although the aviation system had been collected and stored all variety of data, it is often underutilized. This problem can be lead to missed opportunities for improving operations and understand more about the travel behaviour from travellers.

Privacy Concerns: Data breaches and cyberattacks are becoming more likely as more data is gathered, saved, and processed (GeeksforGeeks, 2024). By collecting and extracting detailed data about travel behaviour from travellers will be easily cause privacy concerns. Due to the fact that although by having analysis will gain benefits for aviation industry, but in the point of view of customer and travellers, the airline industry has leaked the user information. Based on this problem, it will let user to more bad impression to the aviation company and will cause the company to have bad review and lesser loyal customer. Therefore, balancing the analysing data with privacy protections is a challenge to the aviation industry.

7.7 Future direction(s) of research in this area.

Future directions of research in travel behaviour analysis in aviation are vital due to the fact that by research and explore more in this part, aviation industry could ensure that the growth and success of the industry, at the same time could let the travellers satisfied and meet the needs of the society. The research included:

Travel Demand Prediction: According to Sharma et al. (2021b), researcher used Travel Behaviour Modelling (TBM) to understand and predict travel decisions by analyse recorded travel information. This research area is crucial to aviation efficiency, capacity management, and optimal flight schedules. Through TBM, better customer experiences, price strategy adjustments, and passenger flow predictions for airports and airlines will be facilitated. By having this research in this area, the benefits will include operational efficiency, revenue optimization, satisfied customers, and improved resource utilization.

Advanced Data Analytics: Advanced Data Analytics offer significant potential to gain deeper insights into passenger behaviour, preferences and popular trends among in aviation research. Advanced data analytics techniques include machine learning and predictive modelling enable airlines can identify patterns and predict future traveller or customer behaviour. Real-time analysis analyses the latest aviation data in no time and change the market conditions, thereby enhancing airline services. Besides that, personalization and customization can analyse the individual behaviour from the past experience and enhance the airline service. Optimization Algorithms can be analysing large volumes of data and generate a most suitable solution to various operational challenges (*Boeing Global Services | Boeing Services*, n.d.).

7.8 Conclusions

In this research, the main purpose is to explore the domain, which was transportation, and with a specific focus in the area of travel behaviour analysis. Based on this research, it could help the aviation industry and related organizations to understand more about the passenger patterns and improve the operation strategies in aviation.

The study identified one of the data mining techniques that are effective in handling complex aviation which was data neural networks (CNN). Additionally, aviation could evaluate the data mining models using methods of cross-validation.

However, there are also challenges that existed in this research. The challenges including data availability and quality and privacy concern by customer. Addressing these challenges is crucial for the successful application of data mining in aviation.

At the same time, there are several directions to improve the data mining problem. These includes travel demand prediction and advanced data analytics. These areas of future research can have more understanding of travel behaviour and lead to substantial advancements in the aviation industry.

In conclusion, the exploration into travel behaviour analysis within the aviation industry underscores the transformative potential of data mining. While challenges such as data quality and privacy concerns exist, the benefits of applying advanced data analytics are substantial. Future research in travel demand prediction and advanced data analytics will pave the way for more refined strategies, ultimately enhancing the efficiency and customer satisfaction in aviation.

Addressing the current challenges and leveraging the identified opportunities can lead to significant advancements. The aviation industry stands to gain from these insights, driving improvements in operational strategies and providing better travel experiences for passengers. This research highlights the critical role of data mining in understanding and optimizing travel behaviour, marking a significant step towards a more efficient and customer-centric aviation industry.

7.9 References

Blackall, M. (2019b, July 1). Global tourism hits record highs - but who goes where on holiday? *The Guardian*. https://www.theguardian.com/news/2019/jul/01/global-tourism-hits-record-highs-but-who-goes-where-on-holiday

Boeing Global Services | Boeing Services. (n.d.). https://services.boeing.com/flight-operations/flight-data-analytics/advanced-safety-analytics

Brownlee, J. (2023, October 4). *A Gentle Introduction to k-fold Cross-Validation*. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/

Convolutional neural Network: benefits, types, and applications. (2023, May 22). Datagen. https://datagen.tech/guides/computer-vision/cnn-convolutional-neural-network/#:~:text=A%20Convolutional%20Neural%20Network%20(CNN,representations%20from%20raw%20input%20images.

Craig, L., & Awati, R. (2024, January 11). *convolutional neural network (CNN)*. Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network

Deepanshi. (2023, August 4). *Beginners Guide to Convolutional Neural Network with Implementation in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/beginners-guide-to-convolutional-neural-network-with-implementation-in-python/#:~:text=The%20Fully%20connected%20layer%20(as,input%20into%20the%20desired%20label.

Dremio. (2023, October 18). *Pooling layers | Dremio*. https://www.dremio.com/wiki/pooling-layers/#:~:text=What%20is%20Pooling%20Layers%3F,retaining%20the%20most%20important%20information.

FareIntelligence. (2023b, November 9). *Dynamic pricing in airlines: The science behind airfare fluctuations*. https://www.linkedin.com/pulse/dynamic-pricing-airlines-science-behind-airfare-fluctuations-pydqf#:~:text=Dynamic%20pricing%2C%20often%20referred%20to,prices%20based%20on%20various%20factors.

Firdausi, M., Ahyudanari, E., & Herijanto, W. (2023). Study on the analysis of travel behavior: A review. *E3S Web of Conferences*, *434*, 02022. https://doi.org/10.1051/e3sconf/202343402022

GeeksforGeeks. (2023a, December 21). *Cross validation in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/cross-validation-machine-learning/

GeeksforGeeks. (2024, April 2). *Challenges of data mining*. GeeksforGeeks. https://www.geeksforgeeks.org/challenges-of-data-mining/

How can you use big data to improve travel behavior analysis? (2023, August 4). www.linkedin.com. https://www.linkedin.com/advice/3/how-can-you-use-big-data-improve-travel#:~:text=For%20example%2C%20you%20can%20use,may%20indicate%20problems%20or%20opportunities.

Melanie. (2024, March 13). *The importance of Cross Validation*. Data Science Courses | DataScientest. https://datascientest.com/en/the-importance-of-cross-validation#:~:text=Cross%2DValidation%20is%20a%20method,to%20work%20on%20new%20data.

Minhas, D. S. (2023, January 7). *Dynamic Airlines Pricing Structure- What we've learned!* https://www.linkedin.com/pulse/dynamic-airlines-pricing-structure-what-weve-learned-singh-minhas#:~:text=Airlines%20use%20complex%20algorithms%20and,decreasing%20ticket%20prices%20as%20needed.

Mishra, M. (2021, December 15). Convolutional neural networks, explained - towards data science. *Medium*. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939#:~:text=Convolution%20Layer&text=It%20carries%20the%20main%20portion,portion%20of%20the%20receptive%20field.

Pankhania, R. (2023, August 16). *Case study – Airline data mining*. https://www.linkedin.com/pulse/case-study-airline-data-mining-riddhi-pankhania#:~:text=Airlines%20use%20data%20mining%20techniques,enhance%20safety%20and%20security%20measures.

Sharma, A., Gani, A., Asirvatham, D., Ahmed, R., Hamzah, M., & Asli, M. F. (2021b). Travel Behavior Modeling: Taxonomy, challenges, and opportunities. *International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications*, *12*(5). https://doi.org/10.14569/ijacsa.2021.0120590

Sevinç, E. (2022, March 1). *An empowered AdaBoost algorithm implementation: A COVID-19 dataset study*.

Computers & Industrial Engineering. https://doi.org/10.1016/j.cie.2021.107912

Simplilearn. (2023, August 10). *What is data mining: definition, benefits, applications, and more*. Simplilearn.com. https://www.simplilearn.com/what-is-data-mining-article#:~:text=Data%20mining%20benefits%20include%3A,profitable%20production%20and%20operational%20adjustments

Tayaba, M., Ayon, E. H., Mia, M. T., Sarkar, M., Ray, R. K., Chowdhury, M. S., Al-Imran, M., Nobe, N., Ghosh, B. P., Islam, M. T., & Puja, A. R. (2023). Transforming customer experience in the airline industry: A comprehensive analysis of Twitter sentiments using machine learning and association rule mining. *Journal of Computer Science and Technology Studies*, *5*(4), 194–202. https://doi.org/10.32996/jcsts.2023.5.4.20

The Editors of Encyclopaedia Britannica. (2024, May 2). *Transportation | Definition & Facts*. Encyclopedia Britannica. https://www.britannica.com/technology/transportation-technology

Types of transportation – RHIHub toolkit. (n.d.). https://www.ruralhealthinfo.org/toolkits/transportation/1/types-of-transportation

Utah Valley University. (2024, January 29). *Technological innovation and disruption*. Pressbooks. https://uen.pressbooks.pub/tech1010/chapter/technological-innovation-and-disruption/

Wolf, C. (2022, September 15). *Aircraft data management: What are the biggest challenges and what can you do — EXSYN*. EXSYN Aviation Solutions. https://www.exsyn.com/blog/aircraft-data-management-what-are-the-biggest-challenges-and-what-can-you-do

8.0 Ravin A/L Kanagarajan

8.1 Introduction

Today, data mining is involved not only in science, but also plays a huge role in a large number of business processes, medicine and finance. With the advent of big data, the need for intelligent analysis has increased, as has the understanding that data may contain more and more hidden patterns and patterns. Even big data analysis is not complete without the use of various mining methods. It covers the needs for making more informed decisions and increasing the level of understanding of various processes and phenomena.

Data mining finds its application in a wide variety of fields and has become an integral part of modern information systems. For example, in business, it helps companies better understand their customers and adapt their products and services to their needs. In medicine, data analysis techniques are used to predict diseases and improve treatment plans. In finance, intelligent analysis helps in risk management and the development of investment strategies.

In this essay, I will describe in detail the clustering method with subtypes and application examples. First, let's go deeper into understanding why data mining is needed at all.

8.2 The motivation for using data mining in the area

Let's touch on some of the main reasons for using intelligent analysis:

Extraction of hidden data: Mostly large data sets contain truly important pieces of information that can be overlooked in traditional data analysis. Intelligent analysis allows you to identify patterns and use them in making important decisions. For example, in the retail sector, the analysis of consumer behavior can reveal hidden trends that will help optimize the product range and increase sales.

Forecasting: In accounting, finance, and healthcare systems, mining techniques are widely used to predict the future behavior of data. This allows you to predict future events that could have an undesirable impact on decisions already made using classical analysis. For example, in finance, time series analysis can predict changes in the market, which allows you to take measures in advance to minimize risks.

Efficiency improvement: It is obvious that deep analysis methods can significantly increase the productivity of a company or a research department, improve the quality of decisions made and, as a result, reduce costs and detect undesirable weaknesses in the current system. In production processes, this can mean optimizing the operation of equipment and reducing downtime, which directly affects the company's profit.

Personalization: Mainly in the field of marketing and direct customer service, intelligent analysis can help to correctly classify customers and, as a result, personalize offers. This allows you to increase customer loyalty and improve the user experience. For example, product recommendations based on previous purchases and customer preferences can significantly increase the likelihood of repeat purchases.

Anomaly detection: Increasing the stability of the system through intelligent anomaly detection can help in the banking sector, as well as in the field of cybersecurity, which also reduces losses both financial and data leaks. For example, real-time transaction analysis can identify suspicious activity and prevent fraud.

Decision support: Intelligent analysis allows the marketing department to make more accurate decisions after the analysis conducted by the analytics department, providing them with the most detailed and understandable dependencies in the data and their other aspects. For example, analyzing market trends and the competitive environment allows you to develop more effective marketing strategies.

Data mining allows you to improve the efficiency and quality of working with data, as well as make more informed decisions. It is used in completely different fields, but the main ones include:

Business: In business, intelligent analysis helps to improve operational processes, optimize supply chains, increase customer satisfaction and increase profits. Companies use data to develop new products, improve customer service, and increase sales.

Healthcare: In healthcare, data analysis helps to improve the quality of medical services, develop personalized treatment plans, predict epidemics, and manage health resources. For example, the analysis of medical records can reveal early signs of diseases, which allows you to start treatment at an earlier stage.

Science: In scientific research, data mining is used to process large amounts of experimental data, identify new patterns, and accelerate the discovery process. For example, in bioinformatics, the analysis of genetic data helps in the development of new treatments and understanding of biological processes.

Education: In the educational field, intellectual analysis helps to improve curricula, adapt learning to the individual needs of students and improve their academic performance. For example, analyzing student performance data allows you to develop personalized educational trajectories and improve teaching methods.

Public administration: In public administration, data analysis is used to improve the quality of services, increase transparency and efficiency of public institutions. For example, analyzing crime data helps to develop more effective crime control strategies and improve public safety.

Data mining plays an important role in the modern world, helping to solve complex problems and improve the quality of life. It allows you to extract hidden knowledge from data, predict the future, improve efficiency, personalize services, detect anomalies and support the decision-making process. In the next section, we will take a detailed look at one of the methods of data mining - clustering.

8.3 The data mining/machine learning techniques used

Clustering:

In data mining, clustering is a process that combines related data items according to their attributes and traits. It is also known as the act of organising a collection of items into groups (called clusters) based on how similar and different the items are to one another. It is an unsupervised learning method designed to find patterns and similarities within a dataset (Jodha, 2023).

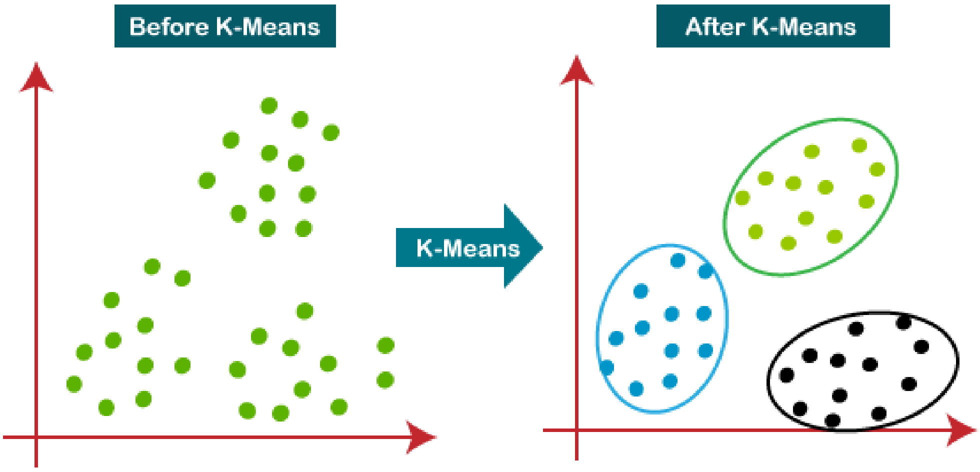


Figure 8.3.2: Example of K-means (K-Means Clustering Algorithm - JavatPoint, n.d.)

One of the most popular and simplest clustering algorithms which was K-means clustering. By using clustering techniques in travel behaviour analysis, clustering can be used to identify patterns and segment the passengers into groups by using similar travel behaviours, can be seen at the Figure 3.3.2. For instance, travellers can be clustered into groups according to the travel habits including budget of travelling, the business travellers, the occasional travellers etc.

The main advantage of using K-means is due to it is simplicity in implementing large dataset (Jain, 2023). Due to the fact that by logically think aviation's system will include a lot of data by customers, by analysing K-means will be simple and easy to implement it by grouping similar data, and afterwards will be easier for analysist to identify pattern the specific part that analysist want to focus analyse, which was travel behaviour analysis in aviation. In the Figure 3.3.1 can be seen the example process of K-means in aviation.

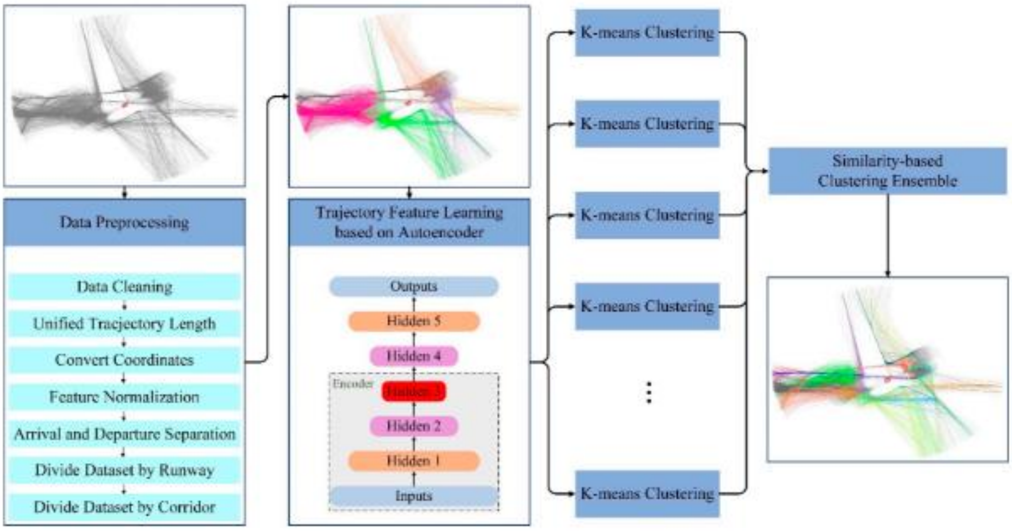


Figure 8.3.1: Example of K-means in aircraft trajectory (Chu et al., 2022)

8.4 The evaluation method of the data mining model

F1 Score mainly is to deal with the confusion matrix and performance metrics. It is a machine learning that measure the model's accuracy and the harmonic mean of precision and recall. By balancing the two indicators, it could do a solitary assessment of a model's effectiveness (Kundu, 2024).

In aviation, they could use F1 score to evaluate classification models including predict the traveller's preferences, booking behaviours or class selection in an airplane. F1 score could generate a detailed understanding of the model and identify the positive and negative cases.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 8.4.1: The formula of F1 Score (Kundu, 2024)

Based on the Figure 8.4.1 can clearly see the formula of F1 Score. In this calculation, the final values will be between 0 and 1, and if the value is 1 it will consist as a superior score (GeeksforGeeks, 2023). The calculation of Precision and Recall will be in the Figure 8.4.2.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Figure 8.4.2: The formula of Precision and Recall (Kundu, 2024)

8.5 New/prospective opportunities produced by the analysis

Operational efficiency improvements in aviation encompass various strategies aimed at optimizing resource utilization, reducing costs, and enhancing overall performance. This includes streamlining flight scheduling and planning to minimize idle time and maximize productivity, optimizing ground operations to reduce turnaround times and delays, and implementing fuel-saving initiatives to minimize environmental impact. Efficient crew management, revenue management, and customer service enhancements are also key components. By focusing on these areas and implementing initiatives to improve operational efficiency, airlines can achieve cost savings, enhance performance, and maintain competitiveness in the industry.

8.6 Challenges faced in this area

Requirements for Skilled Experts: According to Seifert (2006), decision and determination is dependent on the expertise of professionals, and in order for it to be successful, it needs to call for technical and analytical experts who know how to organise the analysis and interpret the results (Revels & Nussbaumer, 2013b). Skilled professionals are essential due to let aviation industry have the accurate and useful conclusions based on the data that been evaluated.

8.7 Future direction(s) of research in this area

Technological Innovation and Disruption: Technological innovation and disruption plays a main role to transform the traditional business models in aviation by implementing new technologies (Utah Valley University, 2024). By improving the technology in aviation could have significantly improve travellers' behaviour, preferences and expectations. The ways of improving technology in aviation including using Artificial Intelligence, Internet of Things (IoT), biometrics, mobile apps etc can have some improvements in safety, efficiency and enhance experience. However, travellers will also have concern in trusting new technologies due to fears of security breaches. To address the concern, aviation industry needs to improve its technology systems while also enhance the security in the system, therefore user and traveller will have trust with aviation industry.

8.8 Conclusion

Clustering plays a key role in modern data mining, offering powerful tools for structuring and understanding large and complex datasets. In the field of air transportation, clustering methods such as k-averages provide unique opportunities for passenger segmentation, which in turn contributes to the creation of personalized marketing strategies and improved service levels.

8.9 References

Chu, X., Tan, X., & Zeng, W. (2022). A clustering ensemble method of aircraft trajectory based on the similarity matrix. *Aerospace*, *9*(5), 269. https://doi.org/10.3390/aerospace9050269

Jain, R. (2023, September 11). K-Means Clustering: use cases, advantages, working principle. *Bombay Softwares*. https://www.bombaysoftwares.com/blog/introduction-to-k-means-clustering

Jodha, R. (2023, June 7). What is Clustering in Data Mining? - Scaler Topics. *Scaler Topics*. https://www.scaler.com/topics/data-mining-tutorial/what-is-clustering-in-data-mining/

J. W. Seifert, "Data Mining and Homeland Security An Overview," CRS Report for Congress, (RL31798), January 2006. - References - Scientific Research Publishing. (n.d.-b). https://www.scirp.org/reference/referencespapers?referenceid=777012

K-Means Clustering Algorithm - JavatPoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Kundu, R. (2024, April 10). F1 Score in Machine Learning: Intro & Calculation. *V7*. https://www.v7labs.com/blog/f1-score-guide

Revels, M., & Nussbaumer, H. (2013, August 28). *Data mining and data warehousing in the airline industry*. https://deliverypdf.ssrn.com/delivery.php?ID=302090085025003077072096003124031074002025091068053033075118065105122006122096018109058055107125012039009107026122096100109014055046044039085011124119119013115079019077050036022066020000118115086103090016106067124107121065074107107095083073067103097025&EXT=pdf&INDEX=TRUE

Utah Valley University. (2024, January 29). *Technological innovation and disruption*. Pressbooks. https://uen.pressbooks.pub/tech1010/chapter/technological-innovation-and-disruption/