Table of Contents

Abstract

In this globalised and dynamic era, data warehousing is imperative in all industries including the aviation industry. This paper presented a brief and concise metadata and data exploration of the merged and extracted datasets about flyers and their airline membership details. Four data types were identified among the 26 variables – eight nominal, two ordinal, four interval and nine ratio attributes. Descriptive statistics of this dataset were detailed in this paper as well. Missing values and outliers were found in these datasets. Five evaluation points of the airline data warehouse are proposed – scalability, performance, data integration and quality, data security and flexibility and support for analytics.

Keywords: data warehouse, initial data exploration, metadata, aviation, airline

1.0 Data Warehousing in the Aviation Industry

Undoubtedly, the aviation or airline industry often utilises the information derived from large volumes of data gathered from different sources to better serve their stakeholders including passengers and business partners, prompting the need for data warehousing. Therefore, in this paper, an initial data exploration on customer flight activity and loyalty history dataset will be detailed together with the corresponding metadata, as well as providing evaluation criteria for the airline data warehouse.

2.0 Metadata

In terms of the detailed encyclopedia of data, also better known as metadata, the chosen datasets about customers' flight activities and membership records were developed as one of the airlines in Canada wanted to examine the implementation of airline loyalty promotional campaigns among their local Canadian flight customers in terms of flight activities and program subscriptions. There are two datasets used, both of which consist of but are not limited to the demographics, flight activities as well as the program enrollment and cancellation status of Canadian customers. These IBM Cognos Analytics originated CSV-formatted datasets are shared on Kaggle, a well-known open-source interactive dataset platform, by one of the platform's active users named Agung Pambudi who is open for queries about the datasets via discussion forums and is expected to keep this dataset up to date monthly if any changes need to be implemented, with the latest update being made around early January 2024 (https://www.kaggle.com/datasets/agungpambudi/airline-loyalty-campaign-program-impact-on-flights).

In terms of the customer flight activity dataset, the 10 attributes involved along with the descriptions of the corresponding values will be detailed below:

- Loyalty Number – the one-of-a-kind membership number of customers
- Year – the year of the period
- Month – the month of the period
- Flights Booked – how many flights were reserved by members themselves alone during the period
- Flights with Companions - how many flights were reserved by members with their companions during the period

- Total Flights – The total number of flight reservations made by members regardless of whether with or without companions
- Distance – Total flight distance measured in kilometres during the period
- Points Accumulated – Membership points earned during the period
- Points Redeemed – Membership points used for rewards exchange during the period
- Dollar Cost Points Redeemed – the total value of membership points used by members in Canadian dollars during the period

On the other hand, for the customer loyalty activity dataset, the 16 attributes involved along with the descriptions of the corresponding values will be detailed below:

- Loyalty Number – the one-of-a-kind membership number of customers
- Country – the country where members reside currently
- Province – the province within the country where members reside currently
- City – the city of the province where members reside currently
- Postal Code – the postal code of where members reside currently
- Gender – members consist of male and female
- Education – the current education status of members, which consists of secondary school and below, college, bachelor, master, and doctorate level
- Salary – the annual salary gained by members (in Canadian dollars)
- Marital Status – whether the members are currently singles, engaged in marriage or divorced
- Loyalty Card – membership status which consists of Star, Nova and Aurora
- CLV – abbreviation of Customer Lifetime Value, which is operationally defined as the total amount of expected expenditure made by flight reserving members
- Enrollment Type – the type of loyalty program subscription, which is either standard or 2018 promotional typed
- Enrollment Year – the year in which the member is enrolled in the loyalty program
- Enrollment Month – the month in which the member is enrolled in the loyalty program
- Cancellation Year – the year in which the member exits the loyalty program
- Cancellation Month – the month in which the member exits the loyalty program

3.0 Initial Data Exploration

3.1 Attribute Types

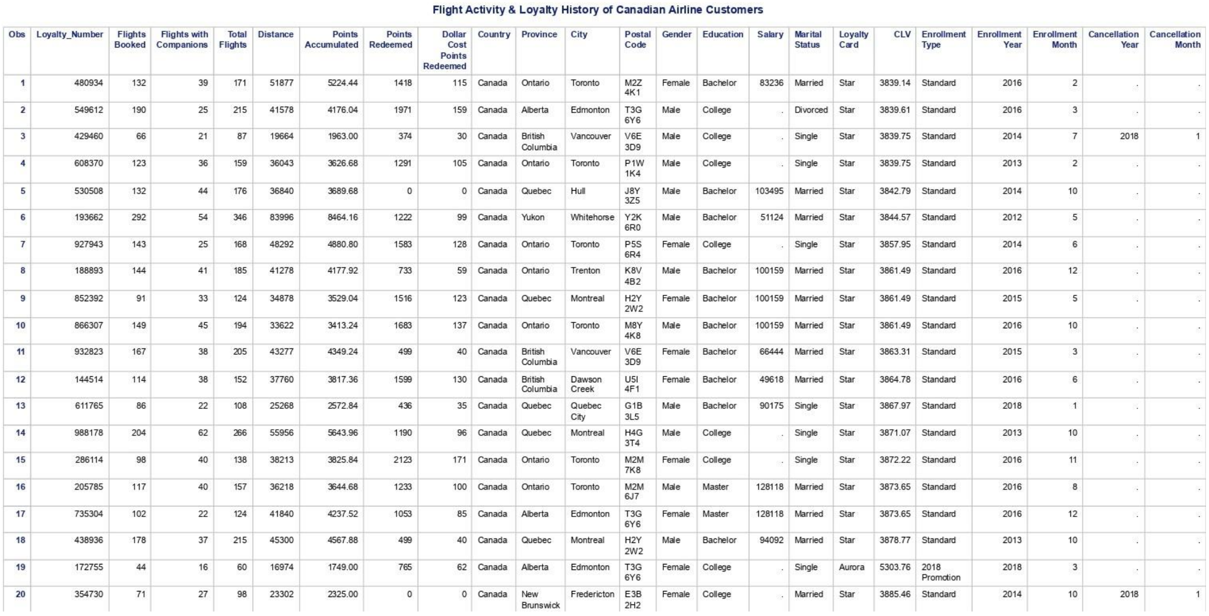
Undoubtedly, initial data exploration is pertinent as one can have a general understanding of the data before proceeding to perform data preprocessing and exploratory data analysis on the chosen dataset accurately. There are four levels of measurement for the involved attributes in the datasets, that are nominal, ordinal, interval, and ratio. For categorical variables, nominal type indicates that the attribute contains two or more levels without any meaningful between-categories rankings, whereas ordinal type is the opposite of nominal type in which the variable also contains two or more levels but with meaningful order between them. Some of the nominal-level variables found in these datasets are loyalty number, country, province, city, postal code, gender, marital status, and enrolment type. Education and loyalty cards are identified as ordinal-level variables because the highest education level is a doctorate, and the highest-ranked loyalty card status is Aurora followed by Nova and Star. On the other hand, the interval type differs from the ratio type in terms of the existence of true zero within the variables. Some of the interval-level attributes found in the two datasets are related to year and month. For ratio, some of the associated variables are flights booked, flights with companions, total flights, distance, points accumulated and redeemed, dollar cost points redeemed, salary and CLV.

3.2 Data Merging and Extraction

The two above-described datasets, customer flight activity and loyalty history, each with 405,624 and 16,737 observations respectively, can be merged into one called ‘Flight_Loyalty’ since both share the same column of loyalty number. To do this, the year and month columns in the flight activity dataset which are non-additive are first removed so that the rest of the numeric and additive columns can be summed with the loyalty number column as the basis due to the relationship between customer loyalty number and other variables being one-to-many. For example, each member can make many flight reservations, but each booking is made by only one member. The first 3,000 observations are then extracted from the merged dataset with 23 variables and 16,737 observations for better computational-efficient data exploration and analysis (see Figures 1 and 2).

Figure 1

An Overview of Flight_Loyalty Dataset (First 20 Observations)

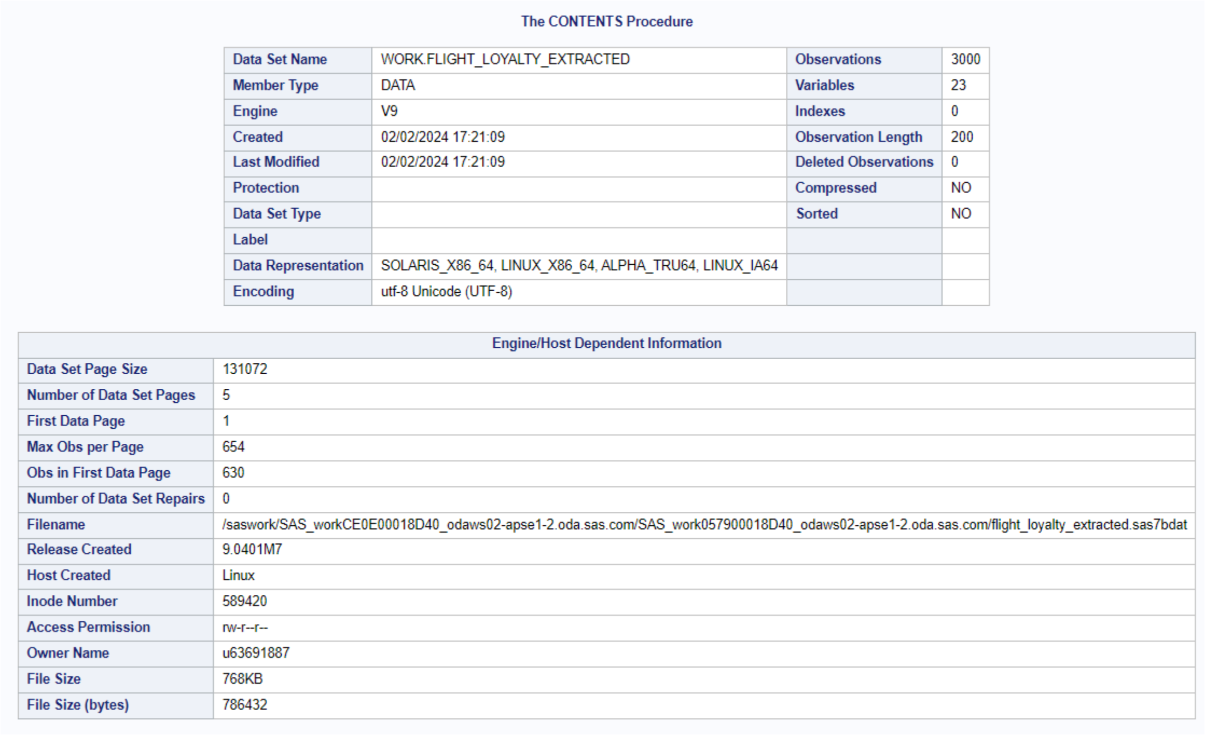


Flight Activity & Loyalty History of Canadian Airline Customers

Obs	Loyalty_Number	Flights Booked	Flights with Companions	Total Flights	Distance	Points Accumulated	Points Redeemed	Dollar Cost Points Redeemed	Country	Province	City	Postal Code	Gender	Education	Salary	Marital Status	Loyalty Card	CLV	Enrollment Type	Enrollment Year	Enrollment Month	Cancellation Year	Cancellation Month
1	480934	132	39	171	51877	5224.44	1418	115	Canada	Ontario	Toronto	M2Z 4K1	Female	Bachelor	83236	Married	Star	3839.14	Standard	2016	2	.	.
2	549612	190	25	215	41578	4176.04	1971	159	Canada	Alberta	Edmonton	T3G 6Y6	Male	College	.	Divorced	Star	3839.61	Standard	2016	3	.	.
3	429460	66	21	87	19664	1963.00	374	30	Canada	British Columbia	Vancouver	V6E 3D9	Male	College	.	Single	Star	3839.75	Standard	2014	7	2018	1
4	608370	123	36	159	36043	3626.68	1291	105	Canada	Ontario	Toronto	P1W 1K4	Male	College	.	Single	Star	3839.75	Standard	2013	2	.	.
5	530508	132	44	176	36840	3689.68	0	0	Canada	Quebec	Hull	J8Y 3Z5	Male	Bachelor	103495	Married	Star	3842.79	Standard	2014	10	.	.
6	193662	292	54	346	83996	8464.16	1222	99	Canada	Yukon	Whitehorse	Y2K 6R0	Male	Bachelor	51124	Married	Star	3844.57	Standard	2012	5	.	.
7	927943	143	25	168	48292	4880.80	1583	128	Canada	Ontario	Toronto	P5S 6R4	Female	College	.	Single	Star	3857.95	Standard	2014	6	.	.
8	188893	144	41	185	41278	4177.92	733	59	Canada	Ontario	Trenton	K8V 4B2	Male	Bachelor	100159	Married	Star	3861.49	Standard	2016	12	.	.
9	852392	91	33	124	34878	3529.04	1516	123	Canada	Quebec	Montreal	H2Y 2W2	Female	Bachelor	100159	Married	Star	3861.49	Standard	2015	5	.	.
10	866307	149	45	194	33622	3413.24	1683	137	Canada	Ontario	Toronto	M8Y 4K8	Male	Bachelor	100159	Married	Star	3861.49	Standard	2016	10	.	.
11	932823	167	38	205	43277	4349.24	499	40	Canada	British Columbia	Vancouver	V6E 3D9	Female	Bachelor	66444	Married	Star	3863.31	Standard	2015	3	.	.
12	144514	114	38	152	37760	3817.36	1599	130	Canada	British Columbia	Dawson Creek	U5I 4F1	Female	Bachelor	49618	Married	Star	3864.78	Standard	2016	6	.	.
13	611765	86	22	108	25268	2572.84	436	35	Canada	Quebec	Quebec City	G1B 3L5	Male	Bachelor	90175	Single	Star	3867.97	Standard	2018	1	.	.
14	988178	204	62	266	55956	5643.96	1190	96	Canada	Quebec	Montreal	H4G 3T4	Male	College	.	Single	Star	3871.07	Standard	2013	10	.	.
15	286114	98	40	138	38213	3825.84	2123	171	Canada	Ontario	Toronto	M2M 7K8	Female	College	.	Single	Star	3872.22	Standard	2016	11	.	.
16	205785	117	40	157	36218	3644.68	1233	100	Canada	Ontario	Toronto	M2M 6J7	Male	Master	128118	Married	Star	3873.65	Standard	2016	8	.	.
17	735304	102	22	124	41840	4237.52	1053	85	Canada	Alberta	Edmonton	T3G 6Y6	Female	Master	128118	Married	Star	3873.65	Standard	2016	12	.	.
18	438936	178	37	215	45300	4567.88	499	40	Canada	Quebec	Montreal	H2Y 2W2	Male	Bachelor	94092	Married	Star	3878.77	Standard	2013	10	.	.
19	172755	44	16	60	16974	1749.00	765	62	Canada	Alberta	Edmonton	T3G 6Y6	Female	College	.	Single	Aurora	5303.76	2018 Promotion	2018	3	.	.
20	354730	71	27	98	23302	2325.00	0	0	Canada	New Brunswick	Fredericton	E3B 2H2	Female	College	.	Married	Star	3885.46	Standard	2014	10	2018	1

Figure 2

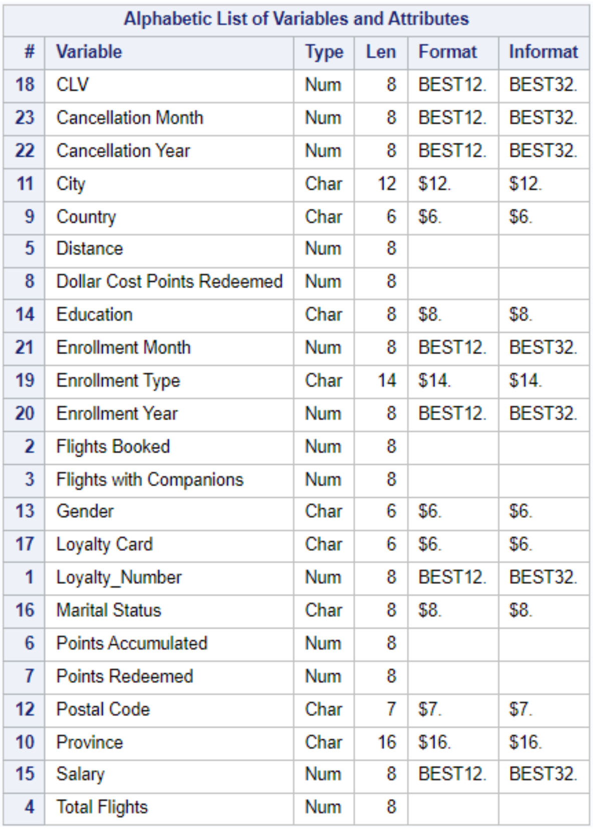
Contents of Flight_Loyalty Dataset



The CONTENTS Procedure

Data Set Name	WORK.FLIGHT_LOYALTY_EXTRACTED	Observations	3000
Member Type	DATA	Variables	23
Engine	V9	Indexes	0
Created	02/02/2024 17:21:09	Observation Length	200
Last Modified	02/02/2024 17:21:09	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	5
First Data Page	1
Max Obs per Page	654
Obs in First Data Page	630
Number of Data Set Repairs	0
Filename	/saswork/SAS_workCE0E00018D40_odaws02-apse1-2.oda.sas.com/SAS_work057900018D40_odaws02-apse1-2.oda.sas.com/flight_loyalty_extracted.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	589420
Access Permission	rw-r--r--
Owner Name	u63691887
File Size	768KB
File Size (bytes)	786432



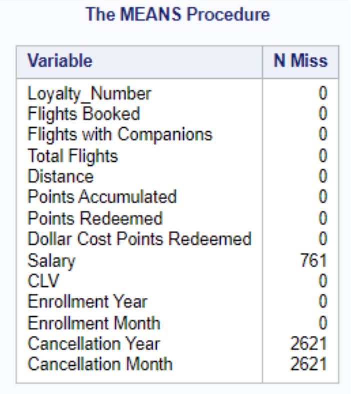
Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
18	CLV	Num	8	BEST12.	BEST32.
23	Cancellation Month	Num	8	BEST12.	BEST32.
22	Cancellation Year	Num	8	BEST12.	BEST32.
11	City	Char	12	$12.	$12.
9	Country	Char	6	$6.	$6.
5	Distance	Num	8		
8	Dollar Cost Points Redeemed	Num	8		
14	Education	Char	8	$8.	$8.
21	Enrollment Month	Num	8	BEST12.	BEST32.
19	Enrollment Type	Char	14	$14.	$14.
20	Enrollment Year	Num	8	BEST12.	BEST32.
2	Flights Booked	Num	8		
3	Flights with Companions	Num	8		
13	Gender	Char	6	$6.	$6.
17	Loyalty Card	Char	6	$6.	$6.
1	Loyalty_Number	Num	8	BEST12.	BEST32.
16	Marital Status	Char	8	$8.	$8.
6	Points Accumulated	Num	8		
7	Points Redeemed	Num	8		
12	Postal Code	Char	7	$7.	$7.
10	Province	Char	16	$16.	$16.
15	Salary	Num	8	BEST12.	BEST32.
4	Total Flights	Num	8		

3.3 Missing Values

In terms of numerical variables, only salary, as well as cancellation year and month columns, have missing values. Specifically, there are 761 missing values for the salary column, whereas there are 2,621 missing values for each cancellation year and month column (see Figure 3). However, there are no missing values in all character variables (see Figure 4). To address possible data inconsistencies, the unique values for each variable are checked and indicated that there are no duplicates among them (see Figure 5).

Figure 3

Missing Values in Flight_Loyalty Dataset (Numerical Attributes)

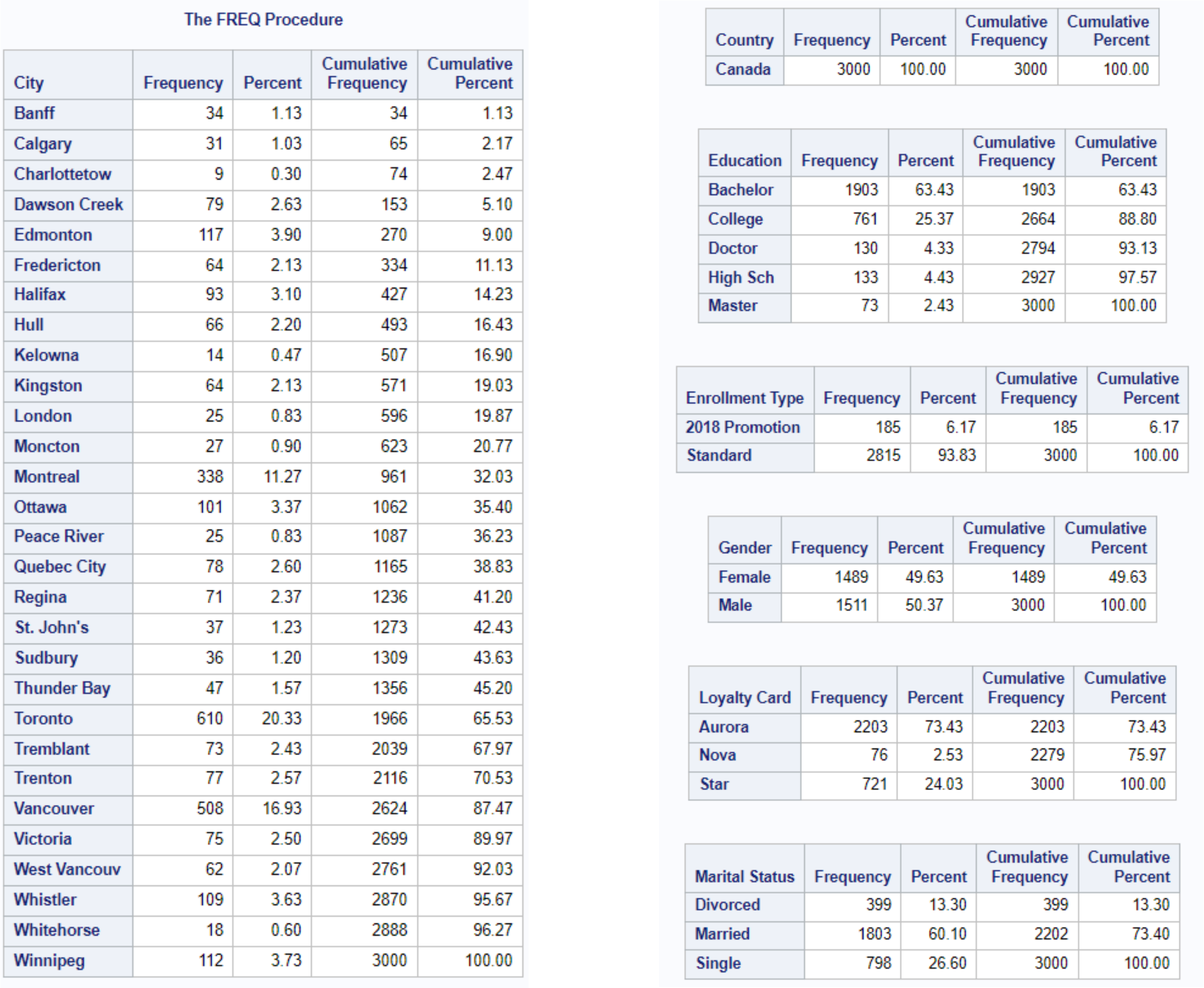


The MEANS Procedure

Variable	N Miss
Loyalty_Number	0
Flights Booked	0
Flights with Companions	0
Total Flights	0
Distance	0
Points Accumulated	0
Points Redeemed	0
Dollar Cost Points Redeemed	0
Salary	761
CLV	0
Enrollment Year	0
Enrollment Month	0
Cancellation Year	2621
Cancellation Month	2621

Figure 4

Missing Values in Flight_Loyalty Dataset (Character Attributes)



The FREQ Procedure

City	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Banff	34	1.13	34	1.13
Calgary	31	1.03	65	2.17
Charlottetow	9	0.30	74	2.47
Dawson Creek	79	2.63	153	5.10
Edmonton	117	3.90	270	9.00
Fredericton	64	2.13	334	11.13
Halifax	93	3.10	427	14.23
Hull	66	2.20	493	16.43
Kelowna	14	0.47	507	16.90
Kingston	64	2.13	571	19.03
London	25	0.83	596	19.87
Moncton	27	0.90	623	20.77
Montreal	338	11.27	961	32.03
Ottawa	101	3.37	1062	35.40
Peace River	25	0.83	1087	36.23
Quebec City	78	2.60	1165	38.83
Regina	71	2.37	1236	41.20
St. John's	37	1.23	1273	42.43
Sudbury	36	1.20	1309	43.63
Thunder Bay	47	1.57	1356	45.20
Toronto	610	20.33	1966	65.53
Tremblant	73	2.43	2039	67.97
Trenton	77	2.57	2116	70.53
Vancouver	508	16.93	2624	87.47
Victoria	75	2.50	2699	89.97
West Vancouv	62	2.07	2761	92.03
Whistler	109	3.63	2870	95.67
Whitehorse	18	0.60	2888	96.27
Winnipeg	112	3.73	3000	100.00

Country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Canada	3000	100.00	3000	100.00

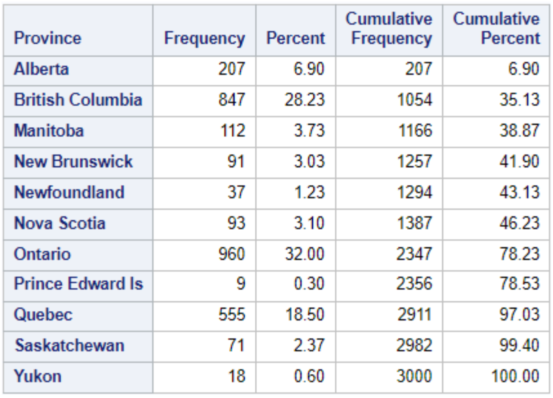
Education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bachelor	1903	63.43	1903	63.43
College	761	25.37	2664	88.80
Doctor	130	4.33	2794	93.13
High Sch	133	4.43	2927	97.57
Master	73	2.43	3000	100.00

Enrollment Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2018 Promotion	185	6.17	185	6.17
Standard	2815	93.83	3000	100.00

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	1489	49.63	1489	49.63
Male	1511	50.37	3000	100.00

Loyalty Card	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Aurora	2203	73.43	2203	73.43
Nova	76	2.53	2279	75.97
Star	721	24.03	3000	100.00

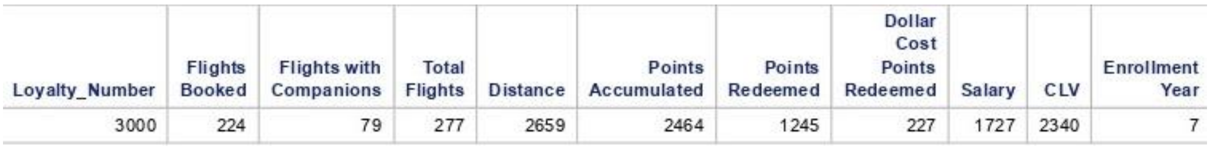
Marital Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Divorced	399	13.30	399	13.30
Married	1803	60.10	2202	73.40
Single	798	26.60	3000	100.00



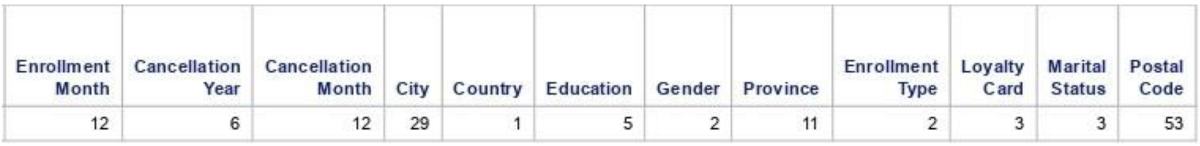
Province	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Alberta	207	6.90	207	6.90
British Columbia	847	28.23	1054	35.13
Manitoba	112	3.73	1166	38.87
New Brunswick	91	3.03	1257	41.90
Newfoundland	37	1.23	1294	43.13
Nova Scotia	93	3.10	1387	46.23
Ontario	960	32.00	2347	78.23
Prince Edward Is	9	0.30	2356	78.53
Quebec	555	18.50	2911	97.03
Saskatchewan	71	2.37	2982	99.40
Yukon	18	0.60	3000	100.00

Figure 5

Unique Values in Flight_Loyalty Dataset



Loyalty_Number	Flights Booked	Flights with Companions	Total Flights	Distance	Points Accumulated	Points Redeemed	Dollar Cost Points Redeemed	Salary	CLV	Enrollment Year
3000	224	79	277	2659	2464	1245	227	1727	2340	7



Enrollment Month	Cancellation Year	Cancellation Month	City	Country	Education	Gender	Province	Enrollment Type	Loyalty Card	Marital Status	Postal Code
12	6	12	29	1	5	2	11	2	3	3	53

3.4 Descriptive Statistics

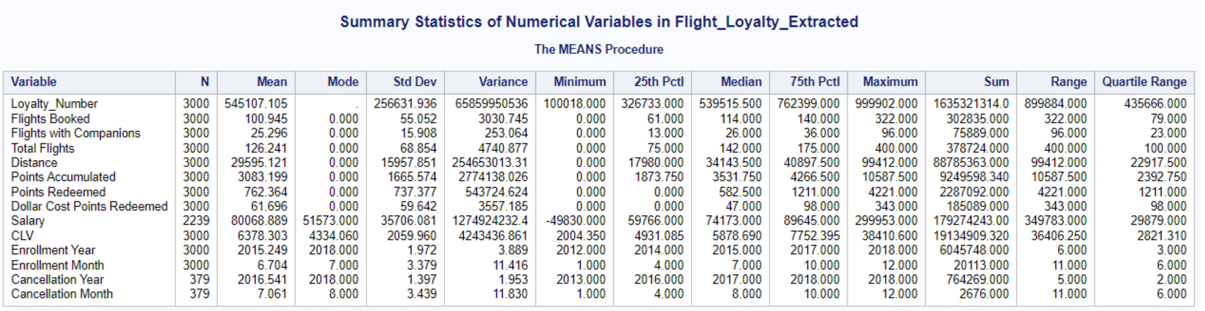
Figure 6 demonstrates the descriptive statistics of the numerical variables in the merged dataset. The average number of flights booked is 101, with a standard deviation of 55 and a variance of 3,031. The minimum and maximum number of flights booked are zero and 322 respectively, while the median number is 114. The number at which 25% and 75% of flights are booked is below 61 and 140 respectively. The average number of flights booked with companions is 25, with a standard deviation of 16 and a variance of 253. The minimum and maximum number of flights booked with companions are zero and 96 respectively, while the median number is 26. The number at which 25% and 75% of flights are booked is below 61 and 140 respectively. The average number of total flights is 126, with a standard deviation of 69 and a variance of 4,741. The minimum and maximum number of total flights booked are zero and 400 respectively, while the median number is 142. The number at which 25% and 75% of total flights is below 75 and 175 respectively. The average number of flight distance travelled is 29,595.12 km, with a standard deviation of 15957.85 km and a variance of 254,653,013.31 km. The minimum and maximum number of distances are zero and 99,412 km respectively, while the median number is 34,143.5 km. The number at which 25% and 75% of the distance travelled is below 17,980 km and 40.897.5 km respectively.

The average points accumulated by members is 3,083.2, with a standard deviation of 1665.574 and a variance of 2,774,138.03. The minimum and maximum points accumulated are zero and 10,587.5 respectively, while the median number is 3531.75. The value at which 25% and 75% of points accumulated is below 1873.75 and 4266.5 respectively. The average points redeemed by members is 762.36, with a standard deviation of 737.38 and a variance of 543724.624. The minimum and maximum points redeemed are zero and 4221 respectively, while the median number is 582.5. The value at which 75% of points are redeemed is below 1211 but there are no points redeemed at the 25th percentile. The average dollar cost points redeemed by members is CDN 61.70, with a standard deviation of CDN 59.64 and a variance of CDN 3557.19. The minimum and maximum dollar cost points redeemed are zero and CDN 343 respectively, while the median dollar is CDN 47. The value at which 75% of dollar cost points are redeemed is below CDN 98 but there are no dollar cost points redeemed at the 25th percentile.

The average annual salary of members is CDN 80,068.89, with a standard deviation of CDN 35,706.081 and a variance of CDN 1,274,924,232.40. The maximum and median annual salary are CDN 299,953 and CDN 74,173 respectively. The value at which 25% and 75% of annual salary is below CDN 59,766 and CDN 89,645 respectively. The average CLV of members is 6,378.3, with a standard deviation of 2,059.96 and a variance of 4,243,436.861. The minimum and maximum CLV are 2,004.35 and 38,410.6 respectively, while the median CLV is 5,878.69. The values at which 25% and 75% of CLV are below 4,931.09 and 7,752.4 respectively. Most airline customers enrolled on and exited the loyalty program in July and August 2018 respectively.

Figure 6

Descriptive Statistics



Summary Statistics of Numerical Variables in Flight_Loyalty_Extracted

The MEANS Procedure

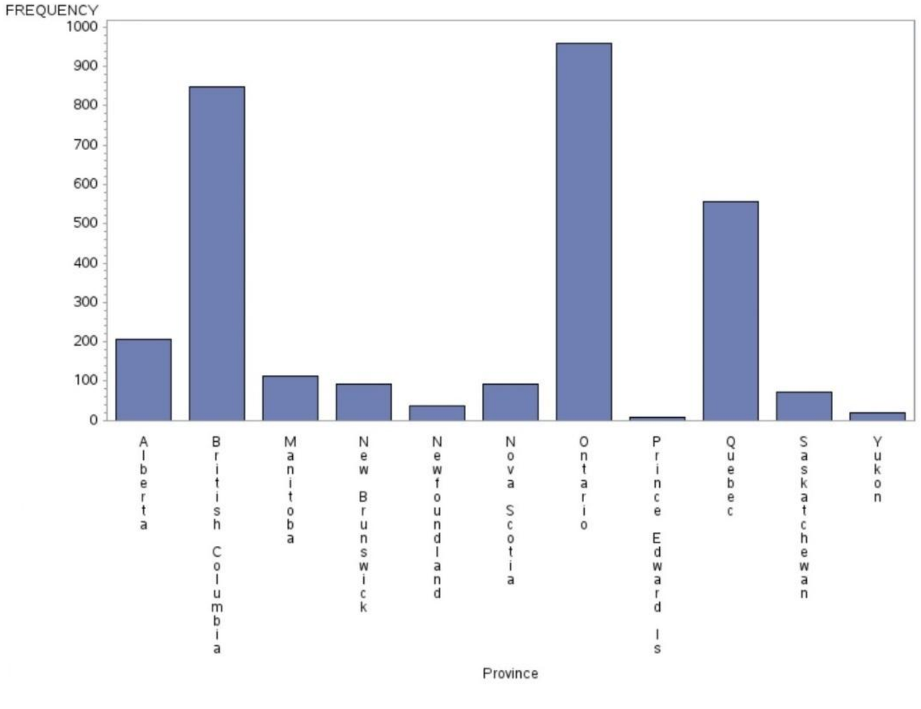
Variable	N	Mean	Mode	Std Dev	Variance	Minimum	25th Pctl	Median	75th Pctl	Maximum	Sum	Range	Quartile Range
Loyalty_Number	3000	545107.105	.	256631.936	65859950536	100018.000	326733.000	539515.500	762399.000	999902.000	1635321314.0	899884.000	435666.000
Flights Booked	3000	100.945	0.000	55.052	3030.745	0.000	61.000	114.000	140.000	322.000	302835.000	322.000	79.000
Flights with Companions	3000	25.296	0.000	15.908	253.064	0.000	13.000	26.000	36.000	96.000	75889.000	96.000	23.000
Total Flights	3000	126.241	0.000	68.854	4740.877	0.000	75.000	142.000	175.000	400.000	378724.000	400.000	100.000
Distance	3000	29595.121	0.000	15957.851	254653013.31	0.000	17980.000	34143.500	40897.500	99412.000	88785363.000	99412.000	22917.500
Points Accumulated	3000	3083.199	0.000	1665.574	2774138.026	0.000	1873.750	3531.750	4266.500	10587.500	9249598.340	10587.500	2392.750
Points Redeemed	3000	762.364	0.000	737.377	543724.624	0.000	0.000	582.500	1211.000	4221.000	2287092.000	4221.000	1211.000
Dollar Cost Points Redeemed	3000	61.696	0.000	59.642	3557.185	0.000	0.000	47.000	98.000	343.000	185089.000	343.000	98.000
Salary	2239	80068.889	51573.000	35706.081	1274924232.4	-49830.000	59766.000	74173.000	89645.000	299953.000	179274243.00	349783.000	29879.000
CLV	3000	6378.303	4334.060	2059.960	4243436.861	2004.350	4931.085	5878.690	7752.395	38410.600	19134909.320	36406.250	2821.310
Enrollment Year	3000	2015.249	2018.000	1.972	3.889	2012.000	2014.000	2015.000	2017.000	2018.000	6045748.000	6.000	3.000
Enrollment Month	3000	6.704	7.000	3.379	11.416	1.000	4.000	7.000	10.000	12.000	20113.000	11.000	6.000
Cancellation Year	379	2016.541	2018.000	1.397	1.953	2013.000	2016.000	2017.000	2018.000	2018.000	764269.000	5.000	2.000
Cancellation Month	379	7.061	8.000	3.439	11.830	1.000	4.000	8.000	10.000	12.000	2676.000	11.000	6.000

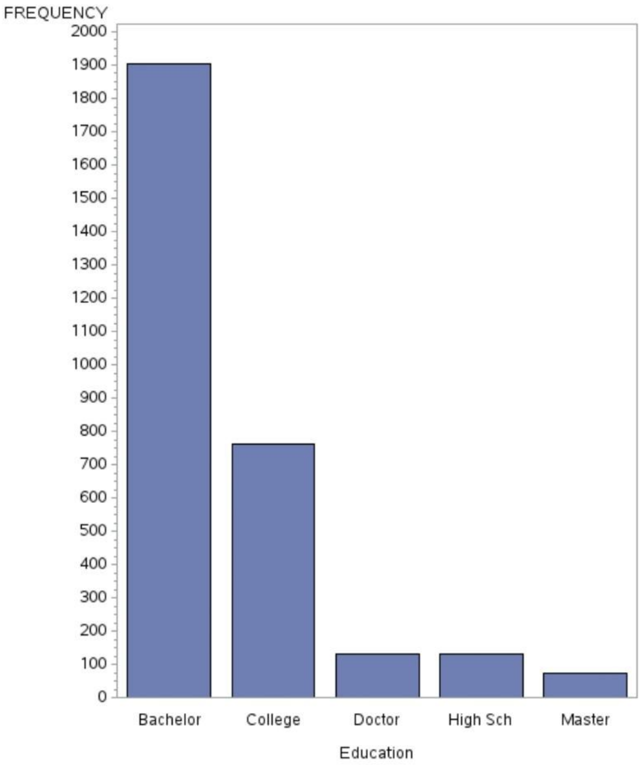
3.5 Frequency of Values

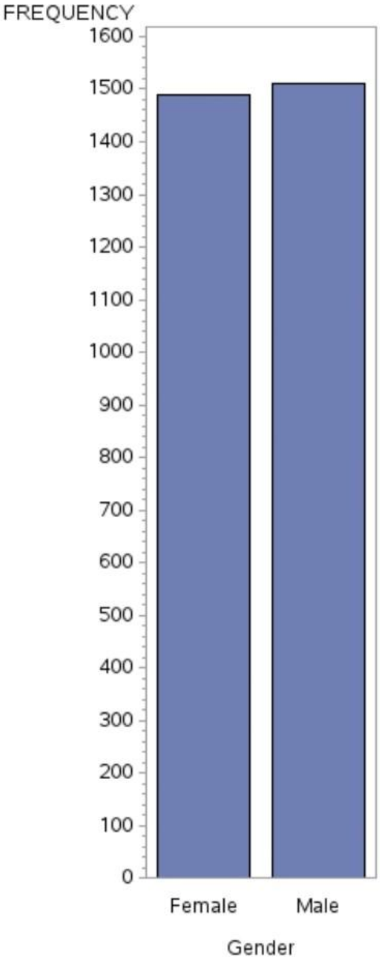
Based on Figure 7, in terms of the frequency of values for character attributes, the city and province where most Canadian airline customers currently reside are Toronto (N = 610) and Ontario (N = 960) respectively. For education level, a majority of 63.43% of the members possess a bachelor’s degree. For enrolment type, 93.83% of them enrolled on the program through the standard pathway, and the rest underwent promotion in 2018. The number of male and female members is quite balanced as indicated by the percentage of males and females which are 50.37% and 49.63% respectively. Among them, 60.1% are married, 26.6% remain single and 13.3% are divorced. For loyalty card type, a majority of 73.43% of the members possess Aurora cards, followed by Star and Nova which are 24.03% and 2.53% respectively.

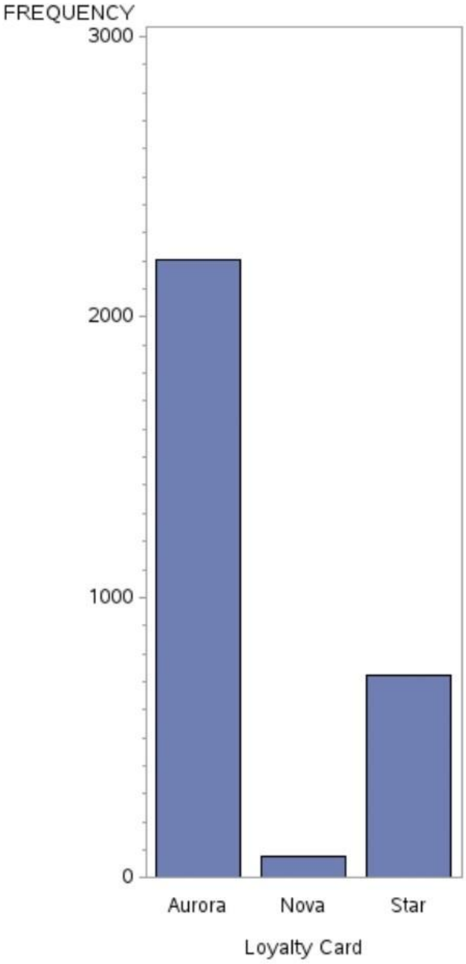
Figure 7

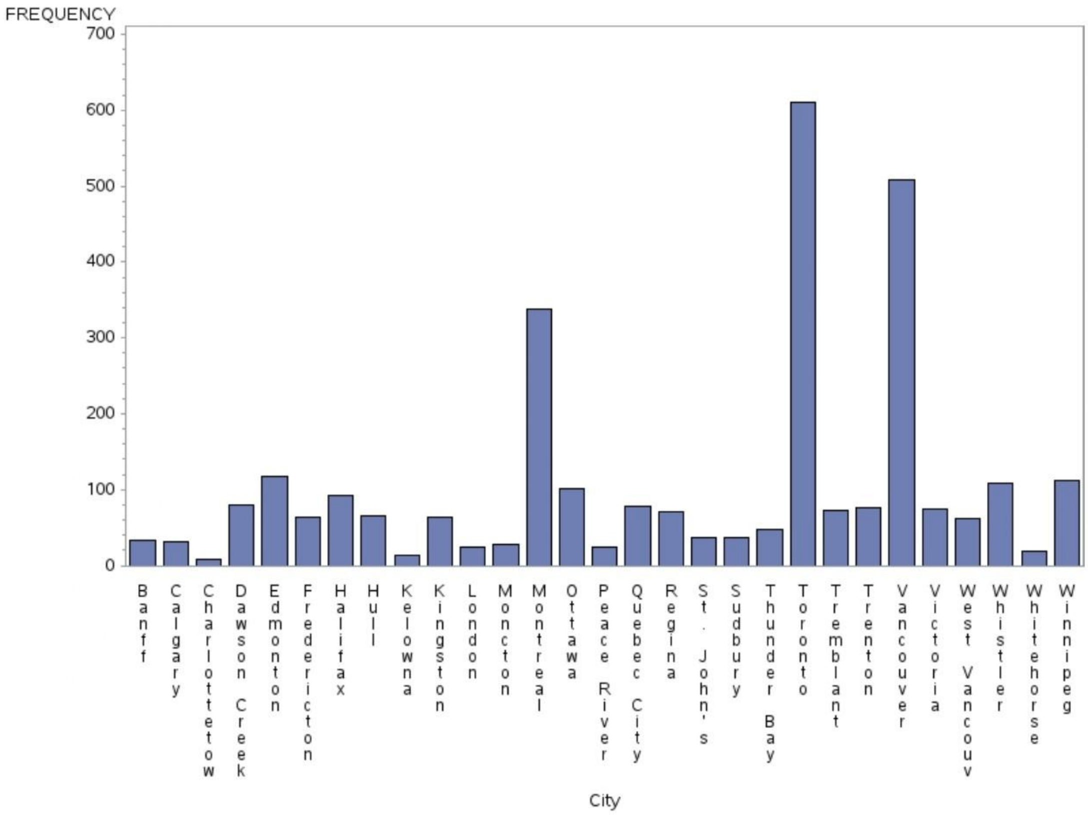
Bar Charts

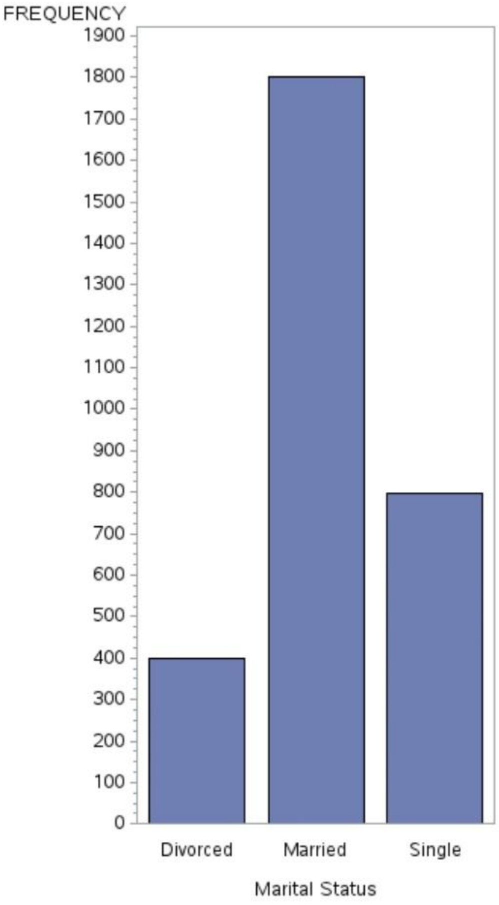


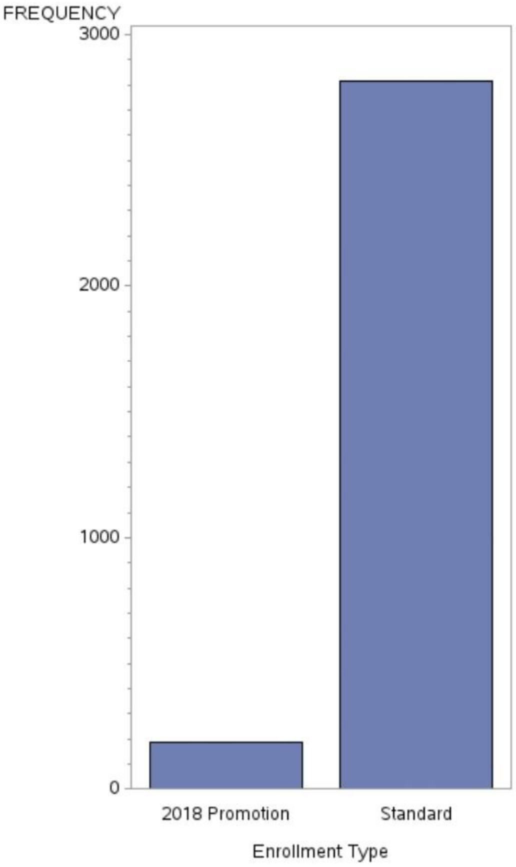

FREQUENCY
2000
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Bachelor
College
Doctor
High Sch
Master
Education


FREQUENCY
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Female
Male
Gender


FREQUENCY
3000
2000
1000
0
Aurora
Nova
Star
Loyalty Card


FREQUENCY
700
600
500
400
300
200
100
0
Banff
Calgary
Charlottetow
Dawson Creek
Edmonton
Fredericton
Halifax
Hull
Kelowna
Kingston
London
Moncton
Montreal
Ottawa
Peace River
Quebec City
Regina
St. John's
Sudbury
Thunder Bay
Toronto
Tremblant
Trenton
Vancouver
Victoria
West Vancouv
Whistler
Whitehorse
Winnipeg
City


FREQUENCY
1900
1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
Divorced
Married
Single
Marital Status

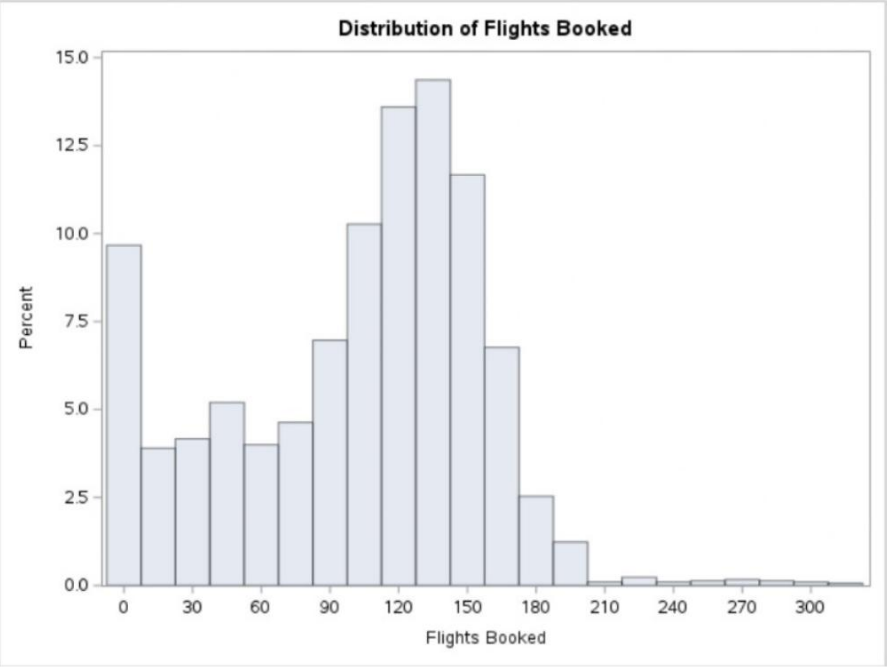

FREQUENCY
3000
2000
1000
0
2018 Promotion
Standard
Enrollment Type

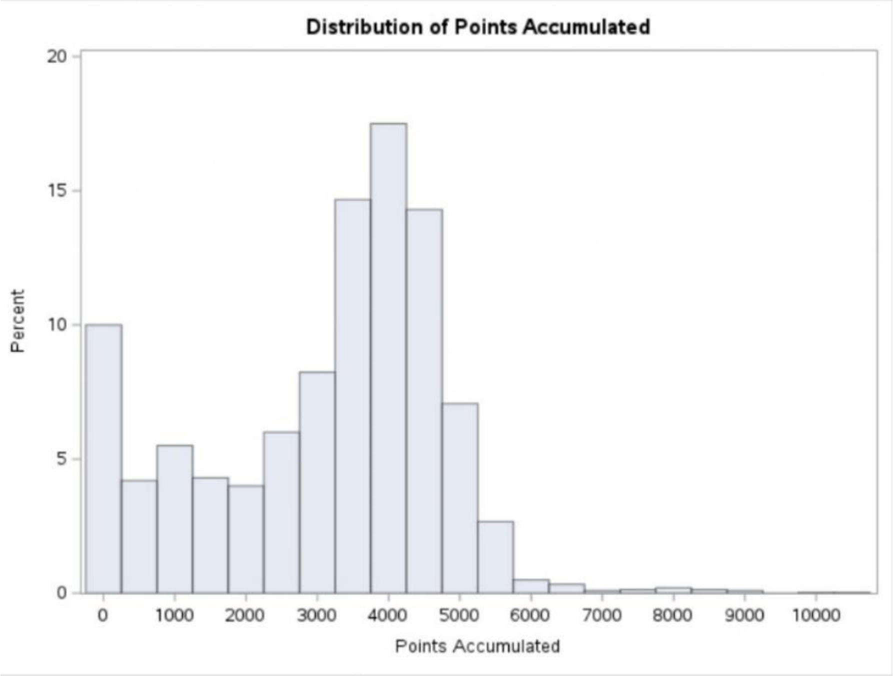
3.6 Histogram

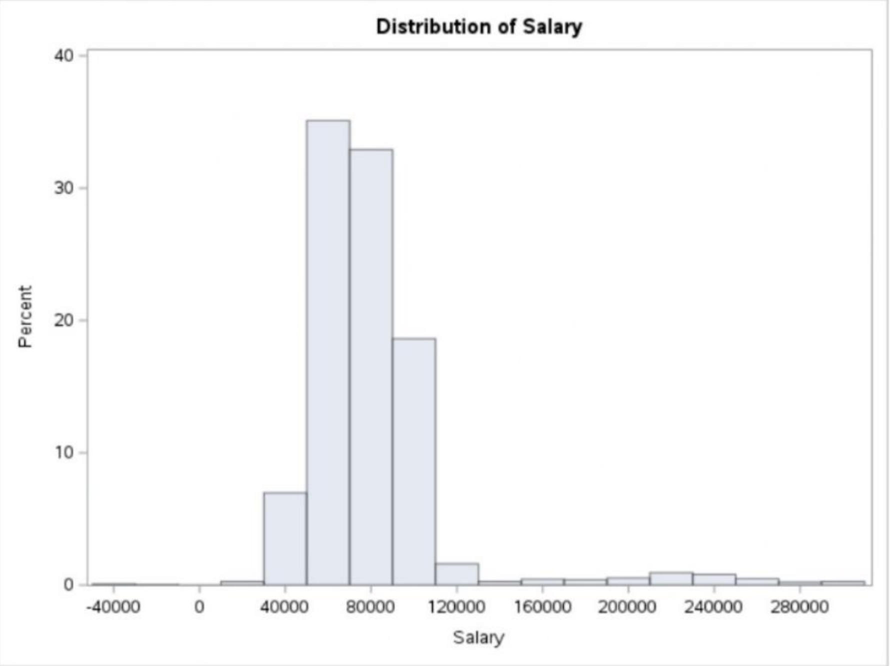
In terms of flight reservations, the histogram distributions of flights booked, flights with companions and total flights are positively skewed based on Figure 8 below. Similarly, the histogram distributions of distance travelled by members as well as their accumulated and redeemed points including dollar cost form, CLV, and annual salary are right-skewed. For time-related attributes, the distributions of cancellation month, as well as enrolment year and month, are approximately symmetrical. However, the distribution of cancellation year is slightly skewed to the positive side.

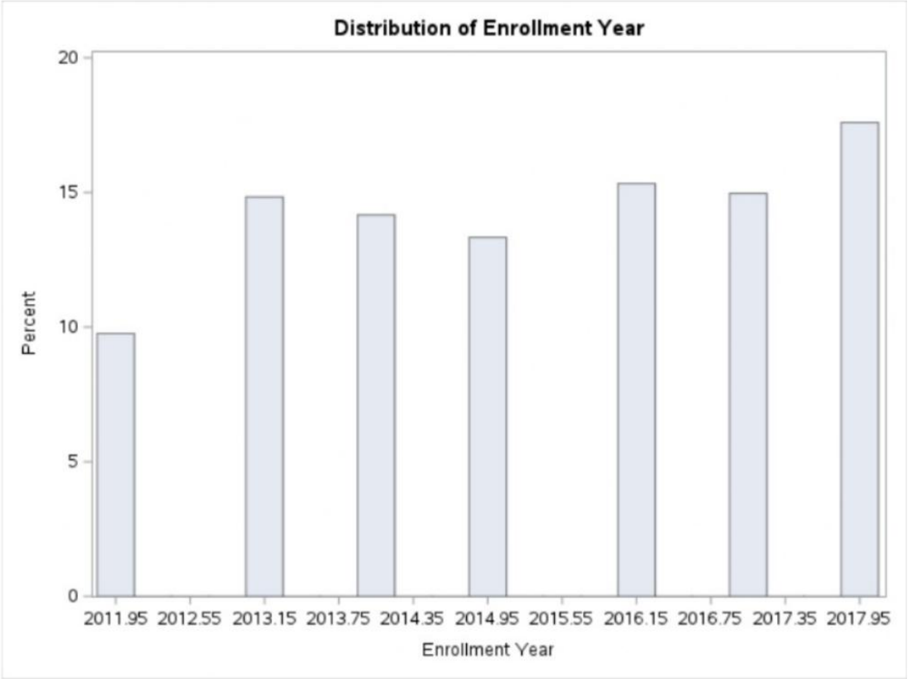
Figure 8

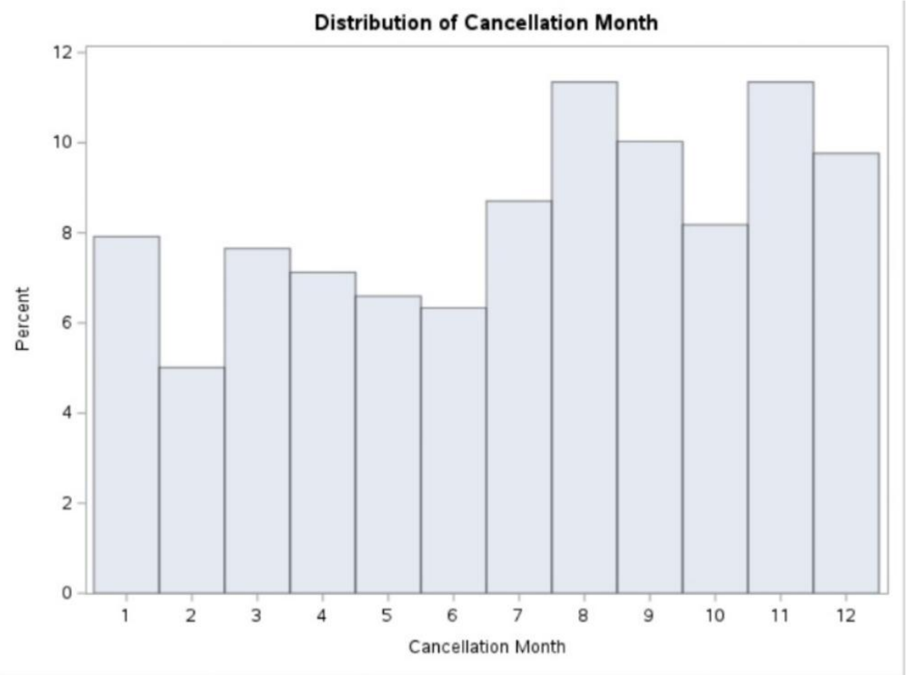
Histograms

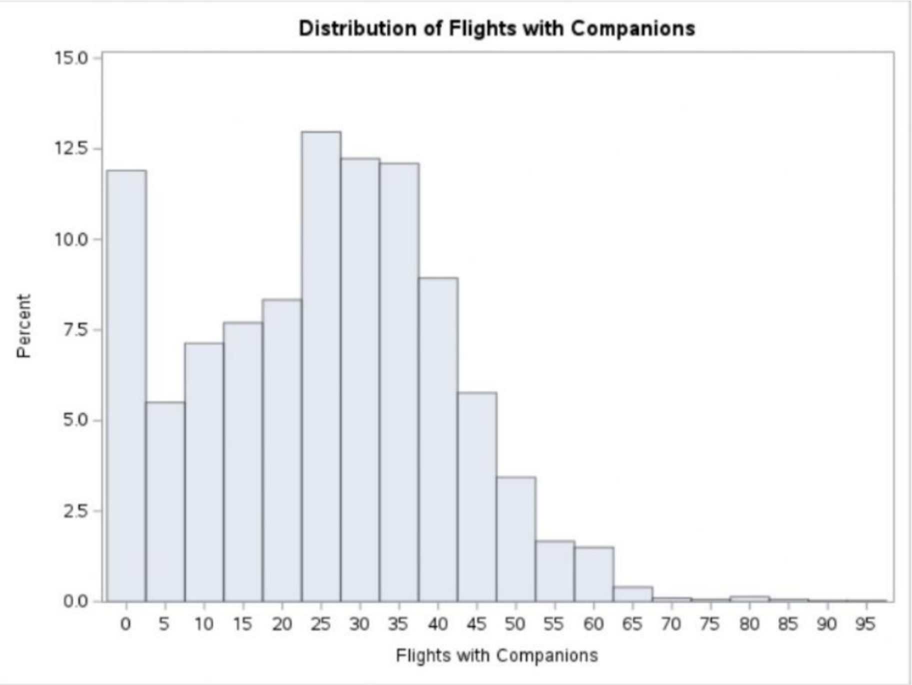


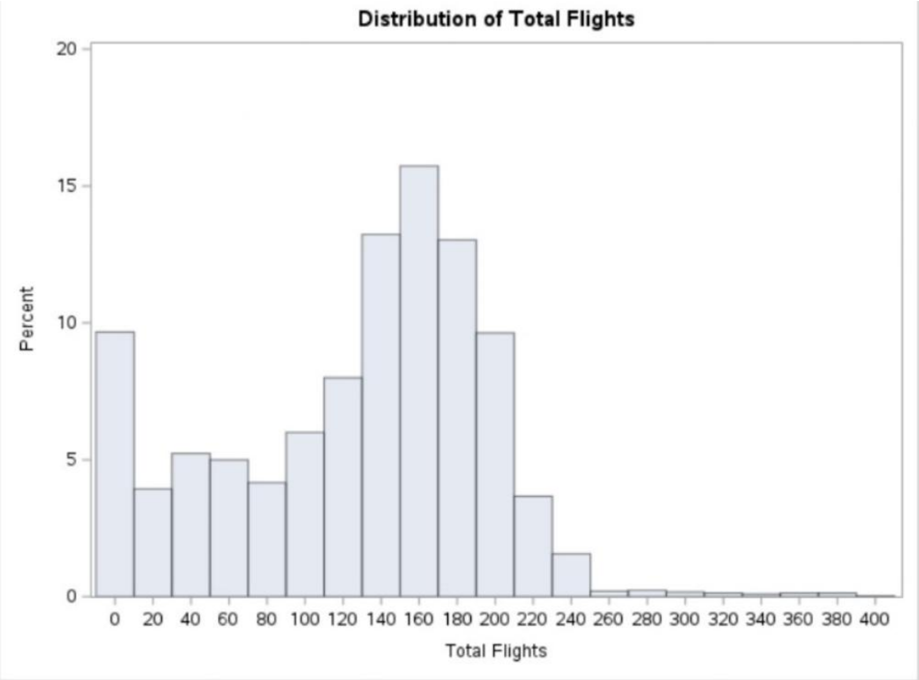

Distribution of Points Accumulated
Percent
0
5
10
15
20
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Points Accumulated

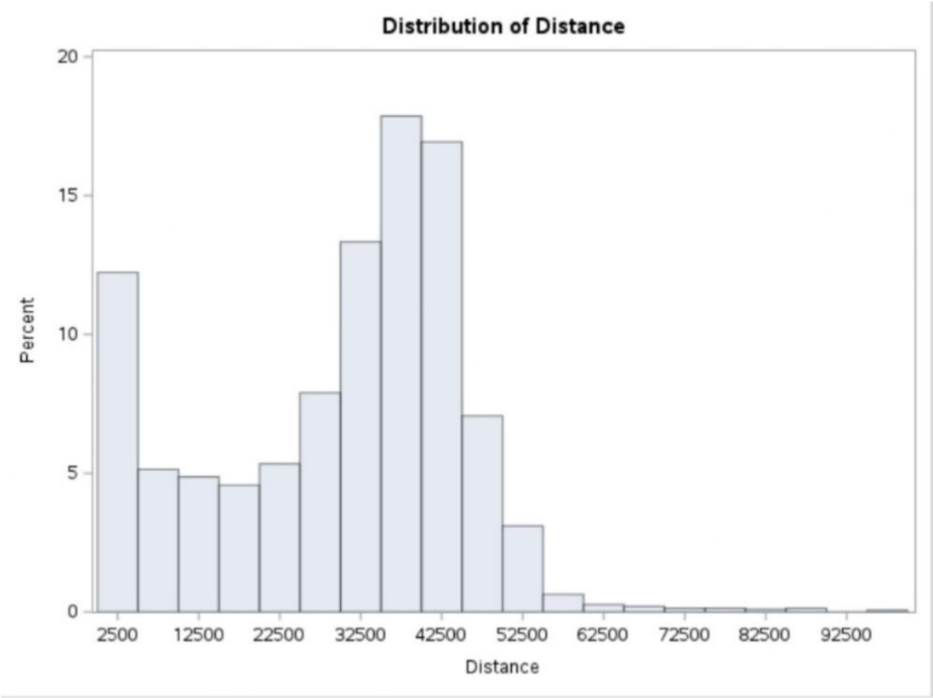

Distribution of Salary
Percent
0
10
20
30
40
-40000
0
40000
80000
120000
160000
200000
240000
280000
Salary

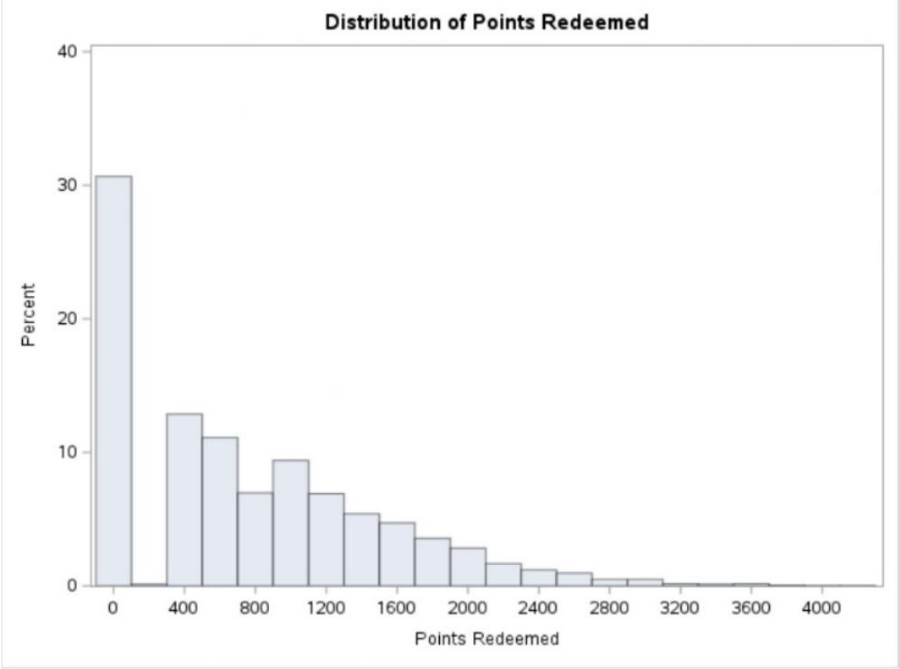

Distribution of Enrollment Year
Percent
0
5
10
15
20
2011.95
2012.55
2013.15
2013.75
2014.35
2014.95
2015.55
2016.15
2016.75
2017.35
2017.95
Enrollment Year

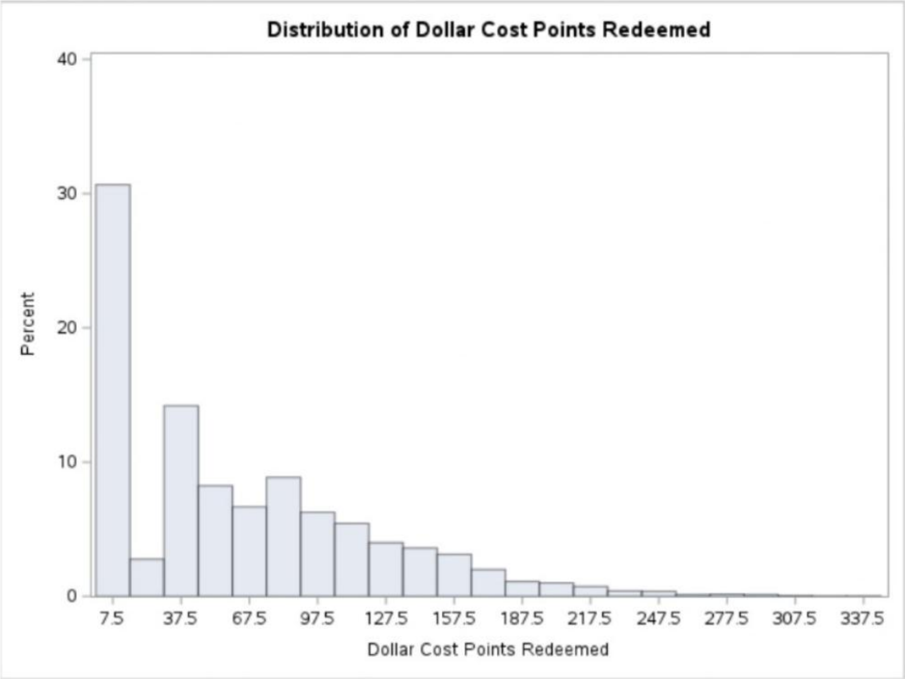

Distribution of Cancellation Month
Percent
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10
11
12
Cancellation Month

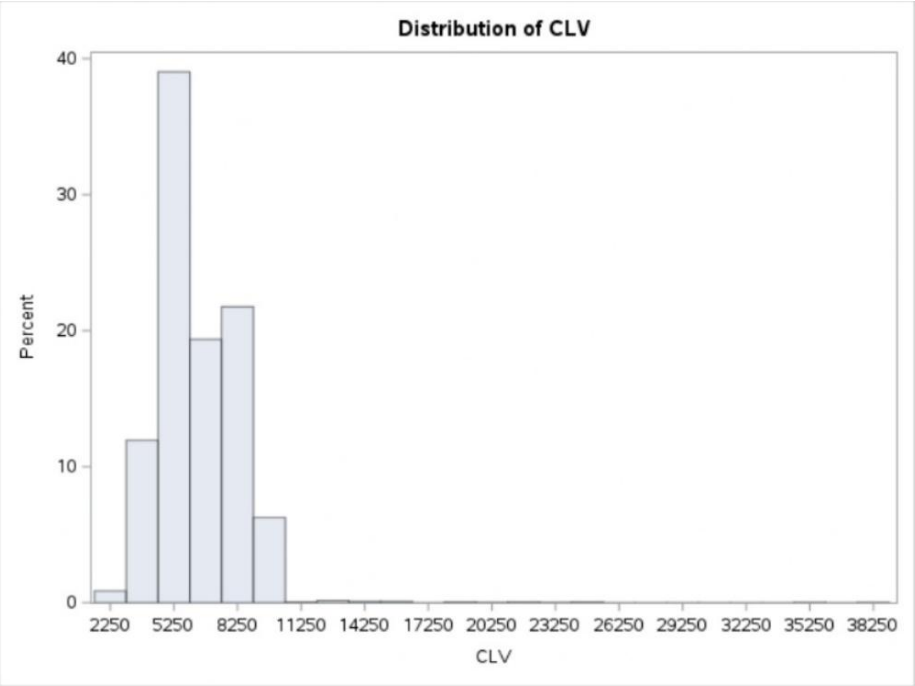

Distribution of Flights with Companions
Percent
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
Flights with Companions

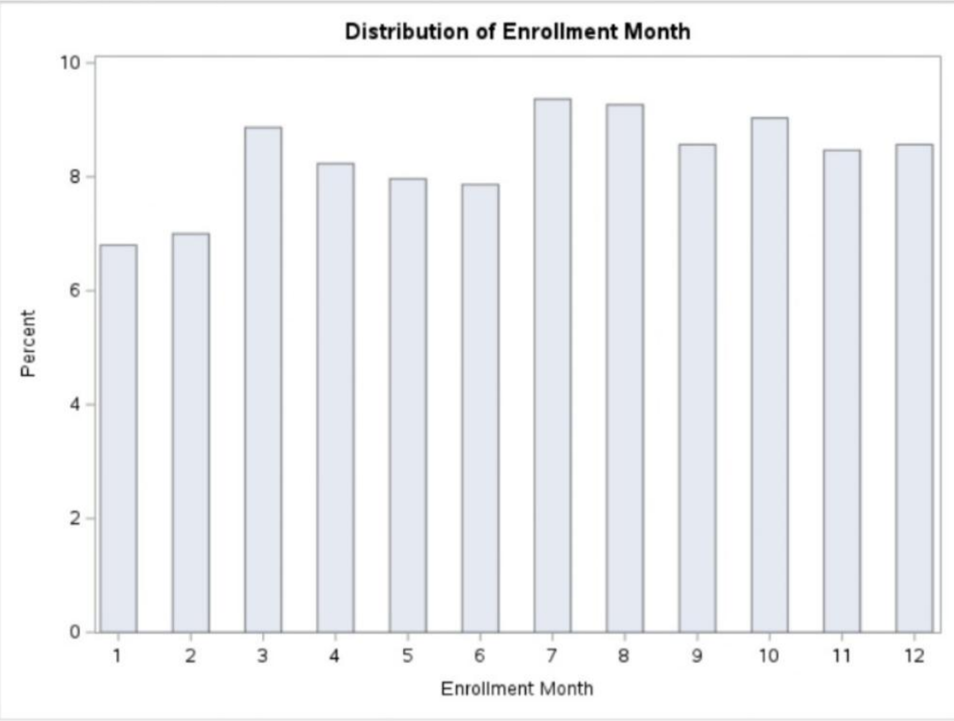

Distribution of Total Flights
Percent
20
15
10
5
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
400
Total Flights

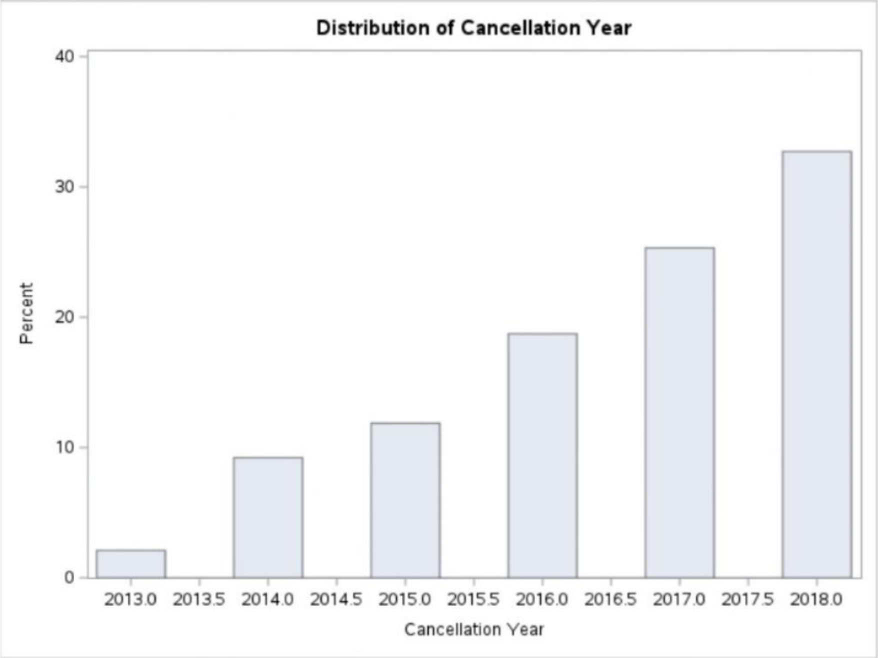

Distribution of Distance
Percent
0
5
10
15
20
2500
12500
22500
32500
42500
52500
62500
72500
82500
92500
Distance


Distribution of Points Redeemed
Percent
0
10
20
30
40
0
400
800
1200
1600
2000
2400
2800
3200
3600
4000
Points Redeemed


Distribution of Dollar Cost Points Redeemed
Percent
40
30
20
10
0
7.5
37.5
67.5
97.5
127.5
157.5
187.5
217.5
247.5
277.5
307.5
337.5
Dollar Cost Points Redeemed


Distribution of CLV
Percent
40
30
20
10
0
2250
5250
8250
11250
14250
17250
20250
23250
26250
29250
32250
35250
38250
CLV



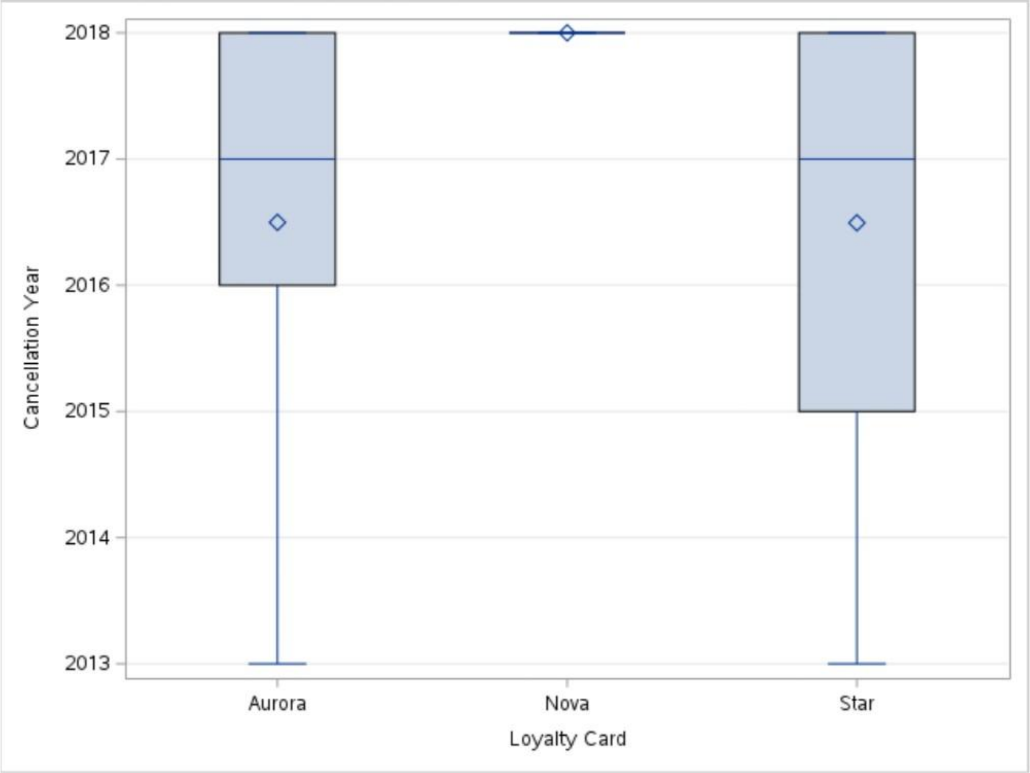


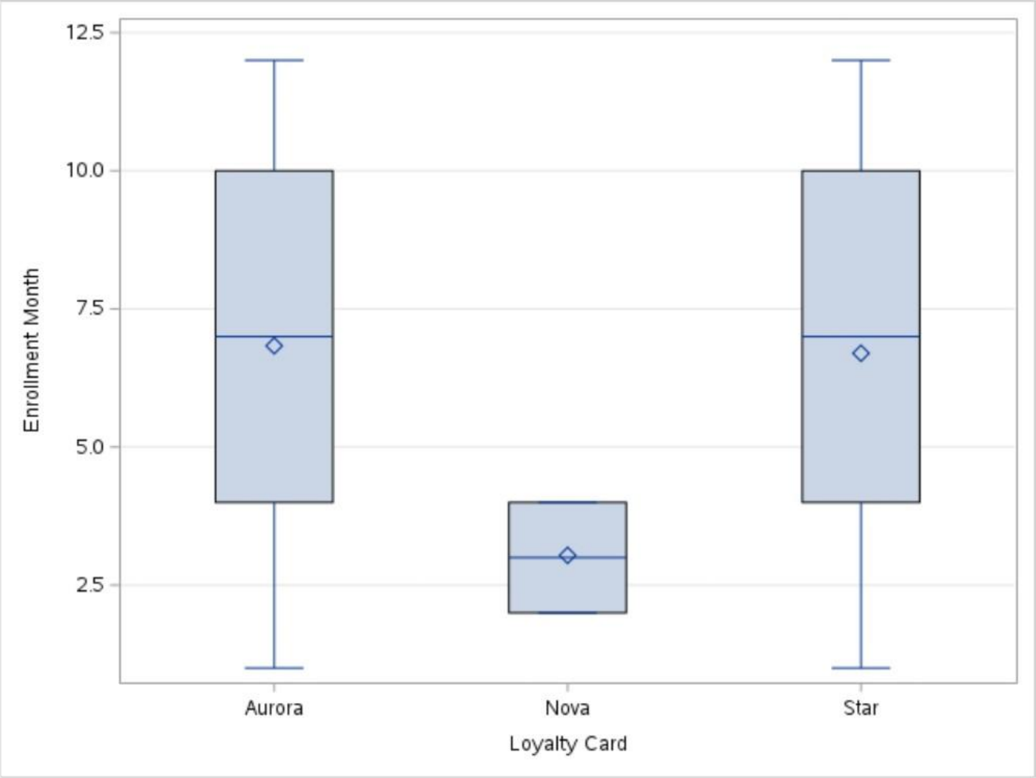
3.7 Boxplot

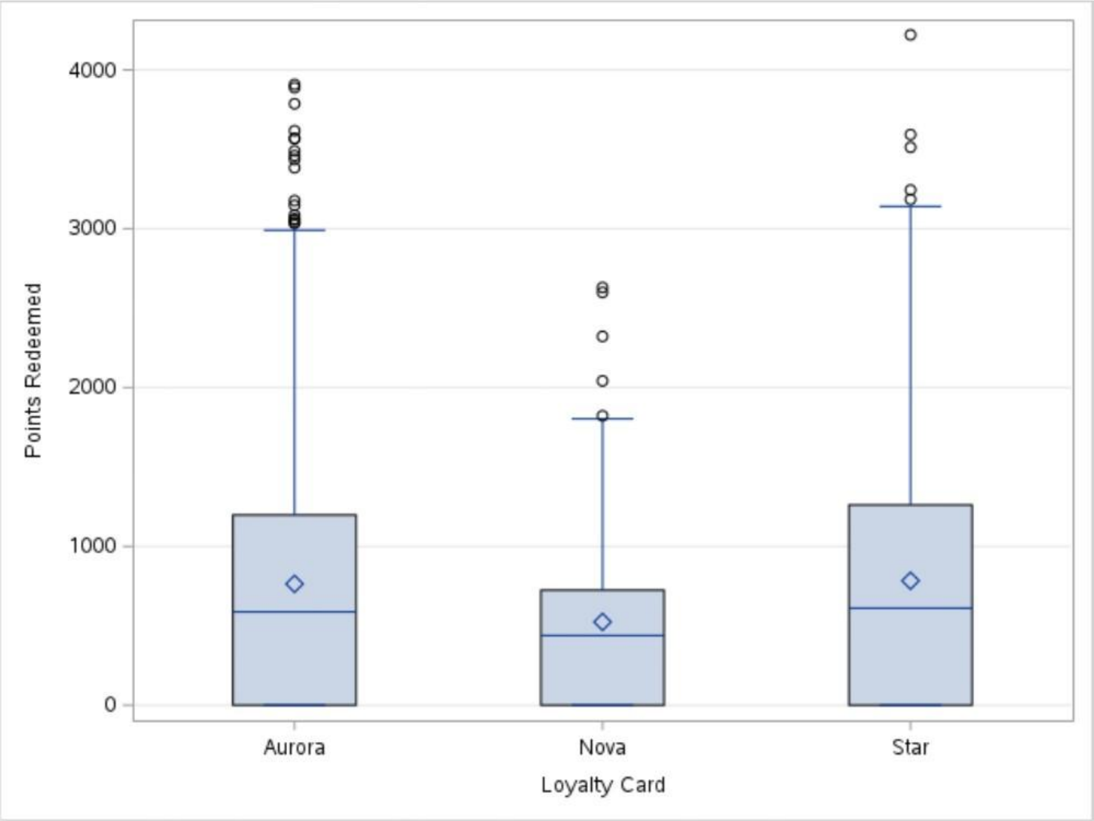
The boxplots below show that in these datasets, there are outliers across all numerical variables on the basis of loyalty card type – Aurora and Star, except for time-related attributes which have no outliers. However, there are no outliers for the Nova loyalty card across numerical variables like distance, flights with companions and points accumulated (see Figure 9).

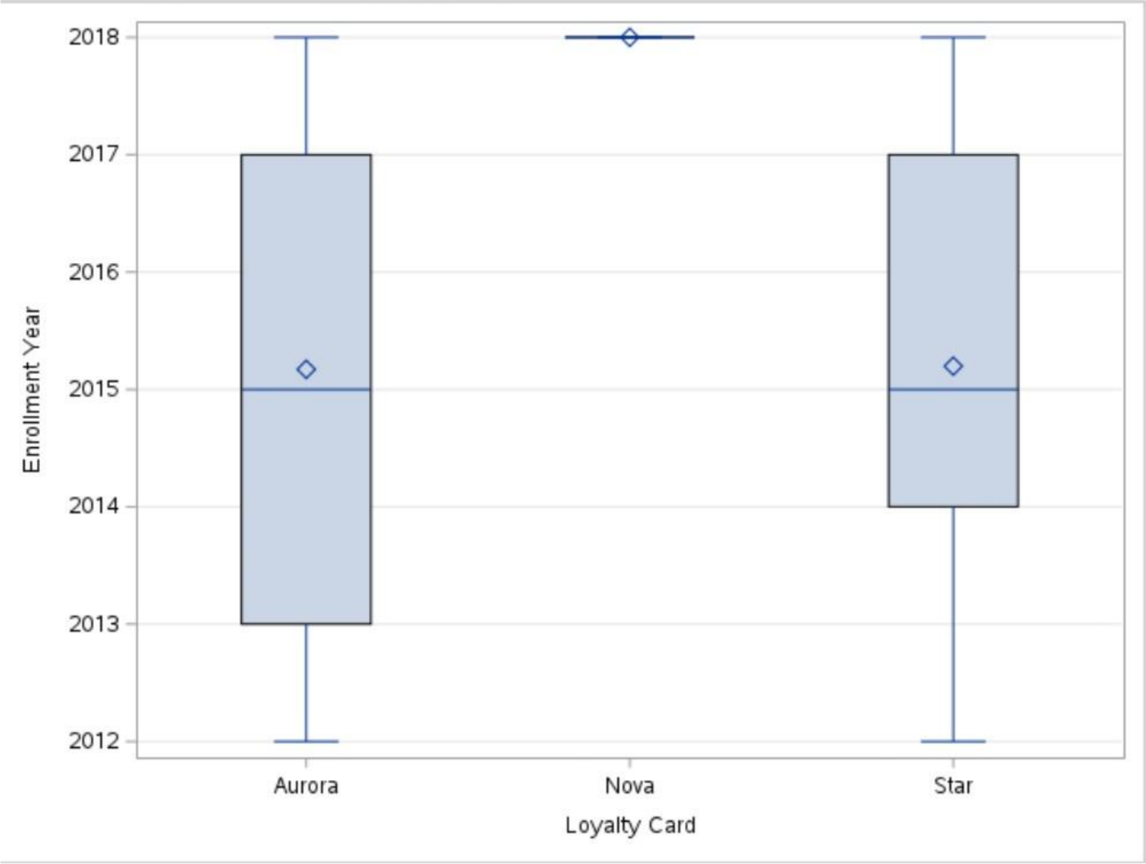
Figure 9

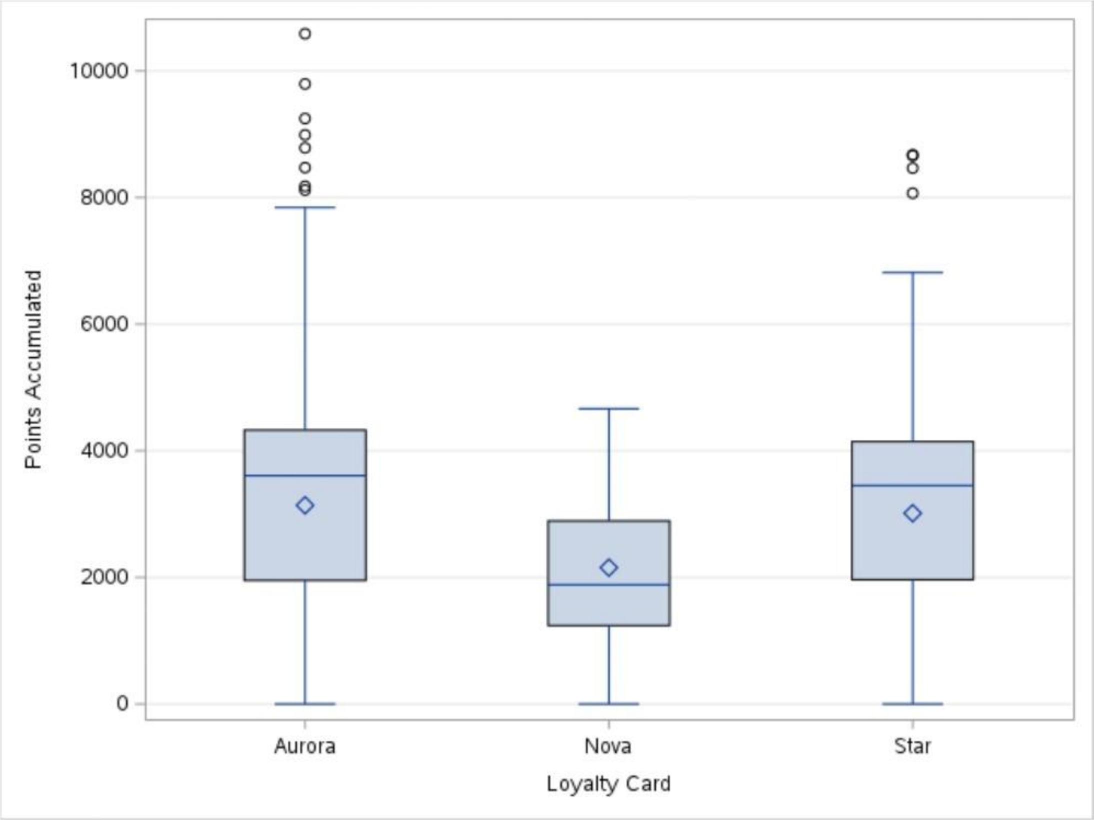
Boxplots



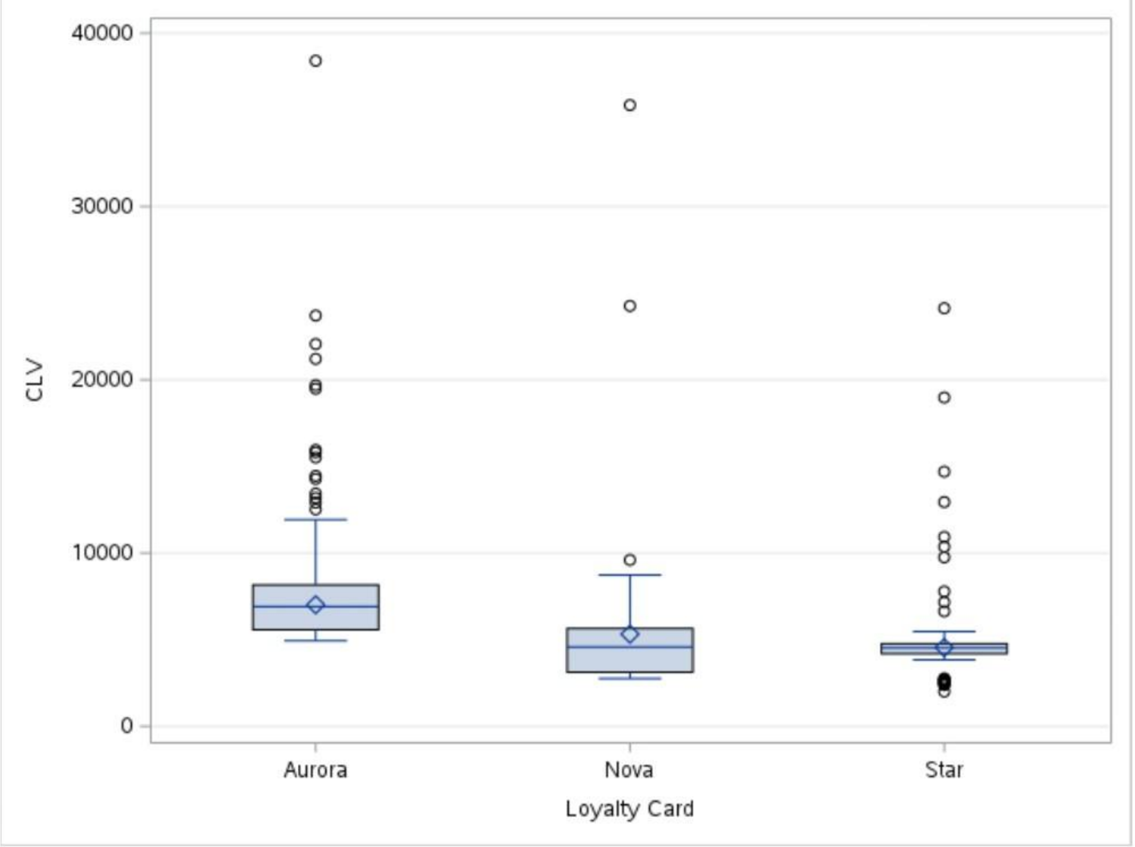



4000
3000
2000
1000
0
Points Redeemed
Aurora
Nova
Star
Loyalty Card

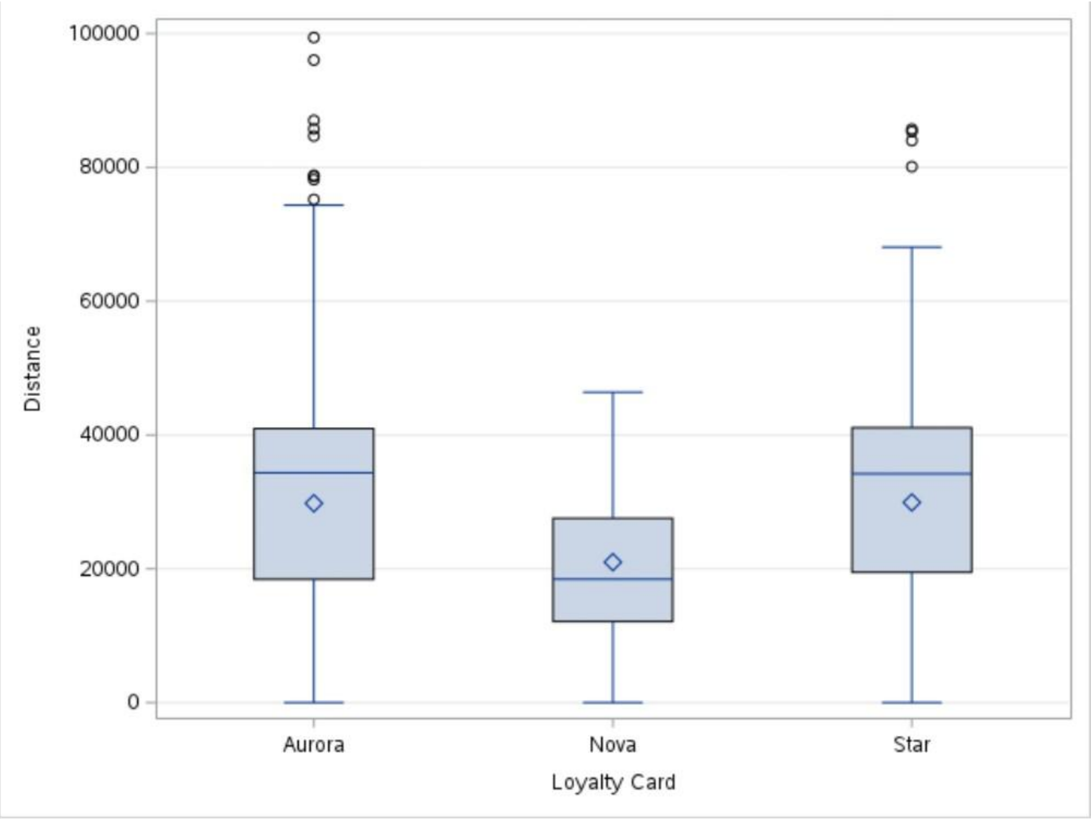

2018
2017
2016
2015
2014
2013
2012
Enrollment Year
Aurora
Nova
Star
Loyalty Card

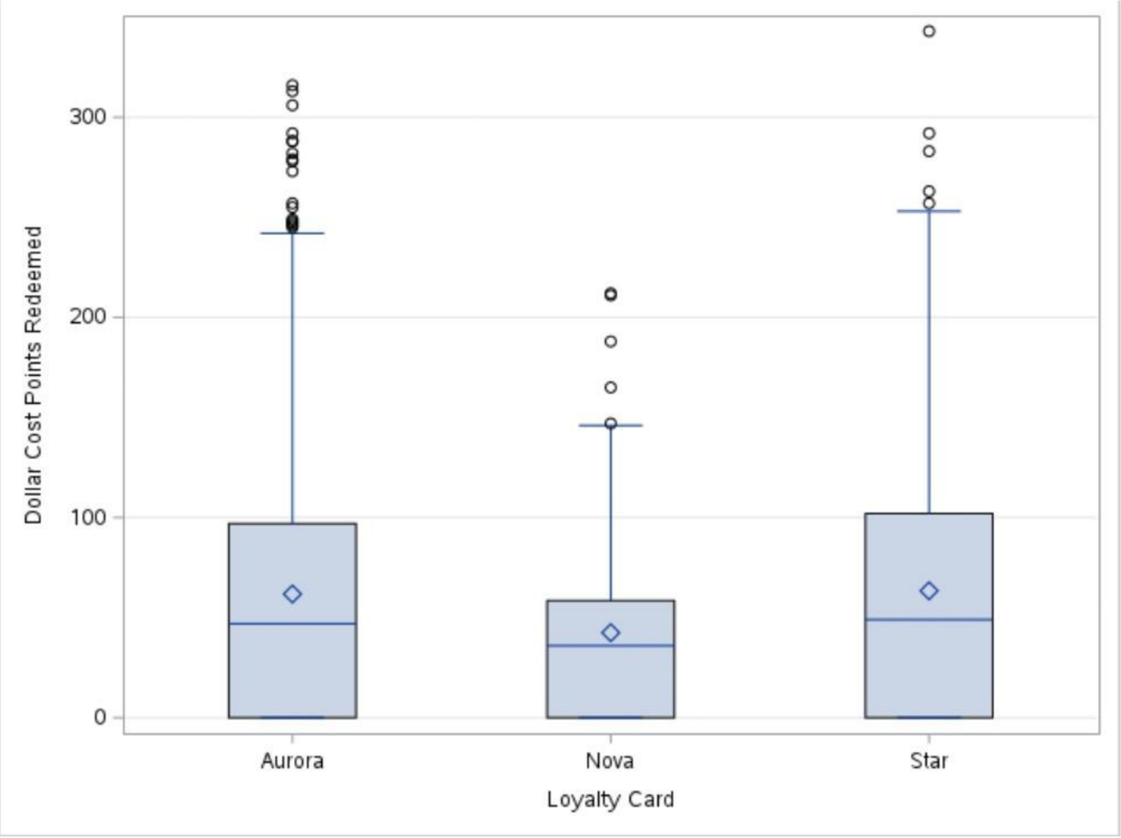


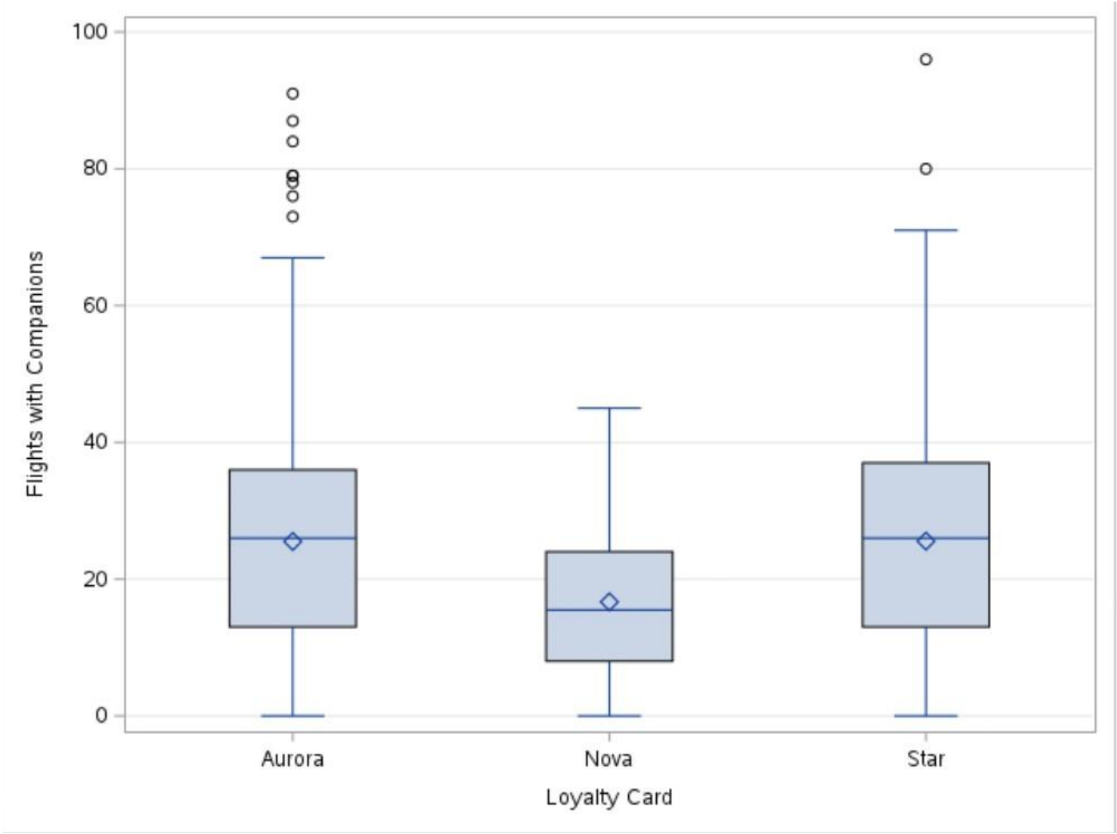
10000
8000
6000
4000
2000
0
Points Accumulated
Aurora
Nova
Star
Loyalty Card

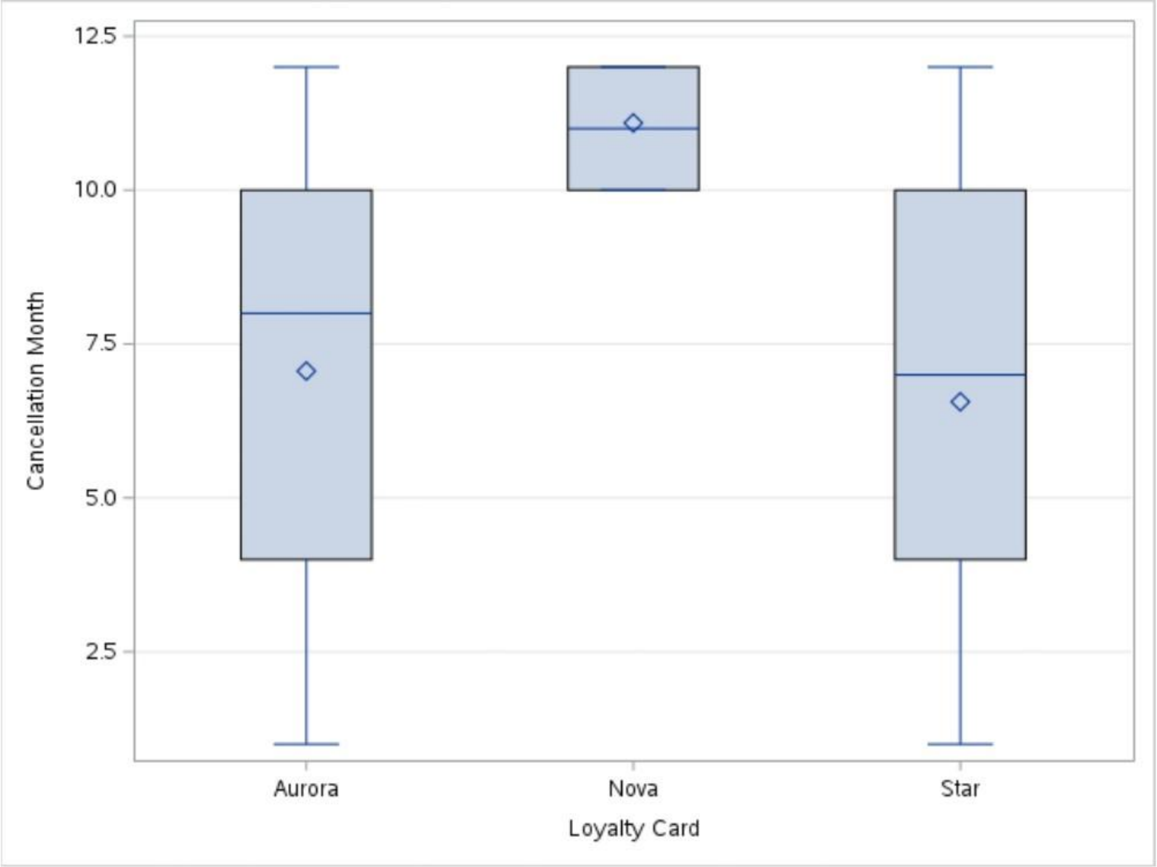


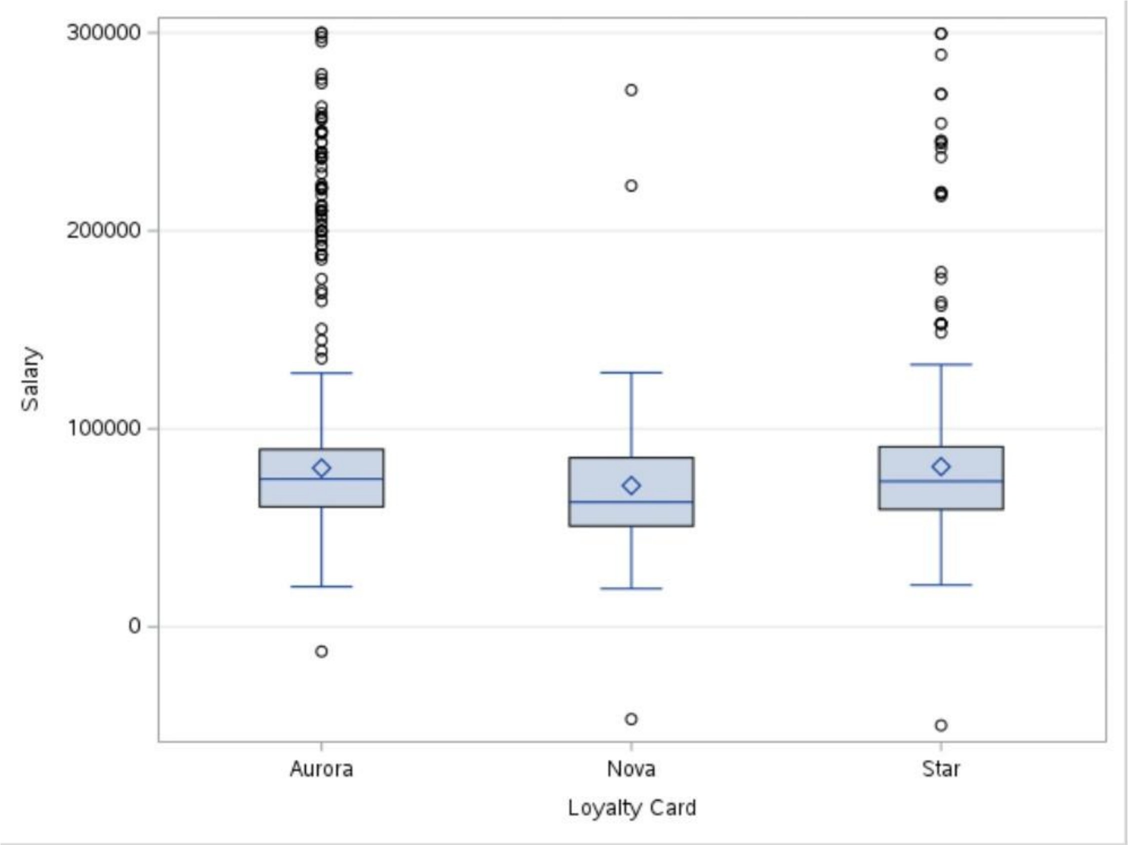
40000
30000
20000
10000
0
CLV
Aurora
Nova
Star
Loyalty Card

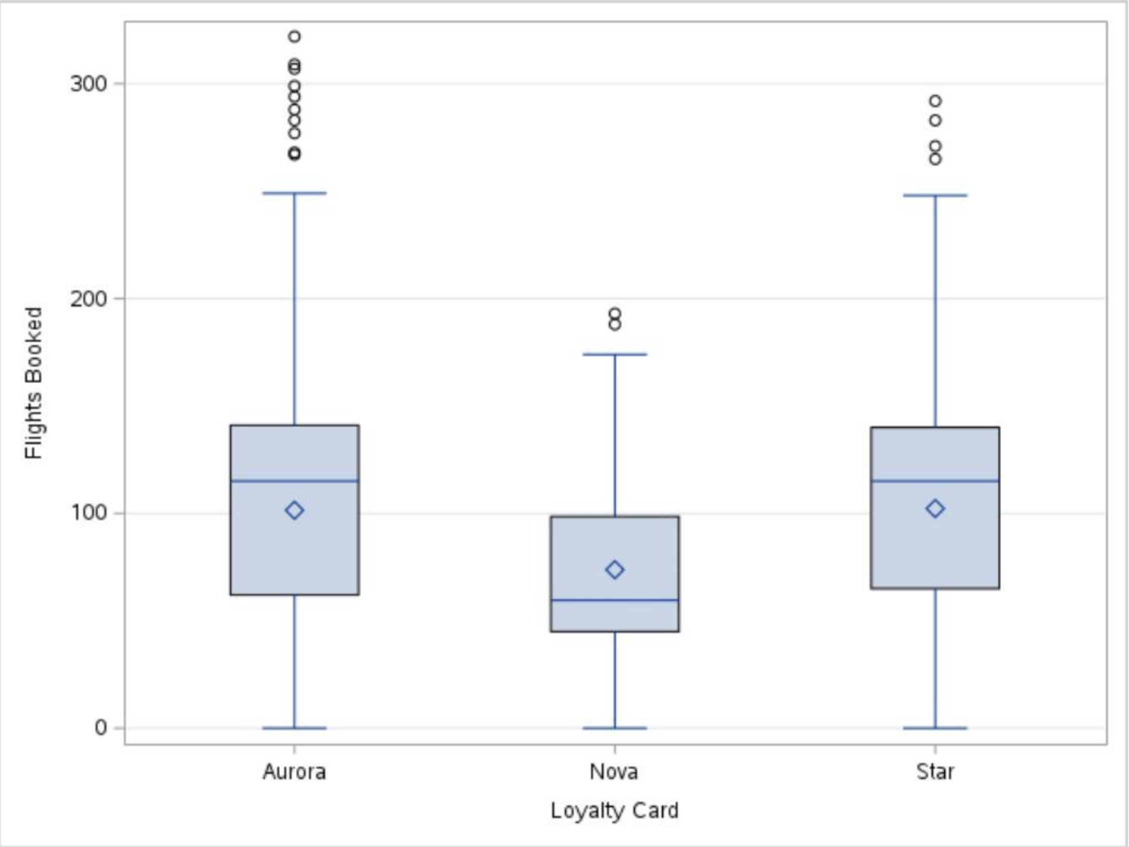

100000
80000
60000
40000
20000
0
Distance
Aurora
Nova
Star
Loyalty Card

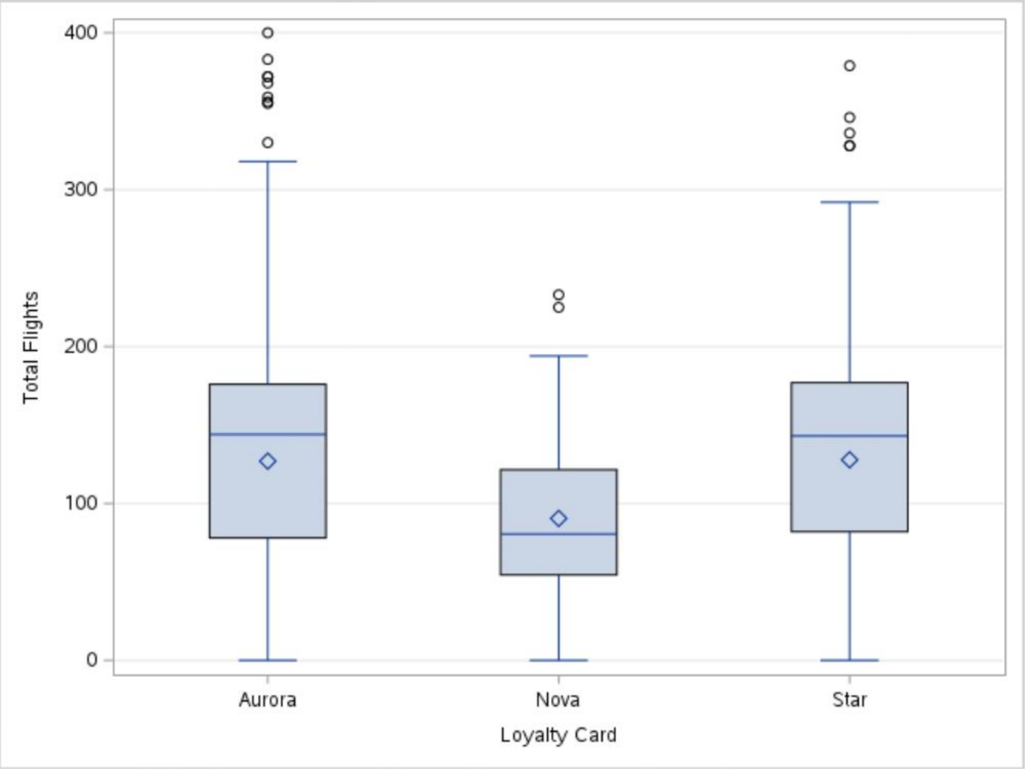

300
200
100
0
Dollar Cost Points Redeemed
Aurora
Nova
Star
Loyalty Card


100
80
60
40
20
0
Flights with Companions
Aurora
Nova
Star
Loyalty Card


12.5
10.0
7.5
5.0
2.5
Cancellation Month
Aurora
Nova
Star
Loyalty Card


300000
200000
100000
0
Salary
Aurora
Nova
Star
Loyalty Card


300
200
100
0
Flights Booked
Aurora
Nova
Star
Loyalty Card



4.0 Evaluate Data Warehouse

The data warehouse is defined as a multidimensional centralised relational database which manages data that are subject-based, time-variant, non-volatile and integrated and it serves as the single source of truth for business stakeholders in the aviation industry to make more informed decisions (Sinha, 2019). It can be evaluated based on five criteria – scalability, performance, data integration and quality, data security and flexibility and support for analytics.

4.1 Scalability

One of the data warehouse evaluation criteria revolves around the scalability aspect. The evolution of the airline company in terms of business landscape can prompt the airline data warehouse to gather increasingly lots of past data from different sources like frequent passenger programs and flight booking systems over time, in addition to the emergence of new major subjects aside from airline passengers like sales of duty-free products at the airport and on the flight and integration of new data sources like airline mobile application (Revels & Nussbaumer, 2013). Therefore, the data warehouse must optimise the scalability of relational databases (up and down depending on business requirements) through distributed parallel processing as a means to keep up with this data growth and evolution so that users can seamlessly query and analyse big data from the scalable data warehouse while maintaining the persistence and integration of the unchangeable and non-updatable historical data from multiple old and new sources (Sinha, 2019). Unlike traditional data warehouses, one of the real-life examples of a highly scalable data warehouse is the implementation of a cloud-based data warehouse which is founded with a massively parallel processing system distributed across relational databases and row-columnar storage (Bani et al., 2018; Rehman et al., 2018).

4.2 Performance

Another data warehouse evaluation criterion is performance. The relationship between scalability and performance of a data warehouse is directly proportional to each other as an increase in scalability leads to a need for an increase in data warehouse performance without time delay which prompts performance monitoring (Antunes et al., 2022; Saddad et al., 2020). This is because as the data warehouse accommodates the need for business growth by the airline company that is associated with an over-time increase of big historical data, it needs to emphasize the personalization of online analytical processing (OLAP) over online transactional processing (OLTP) on the specific airline business needs through targeted fast queries and reports revolving around past data which are related to the particular business subjects within the airline industry so that they can make faster and more informed decisions from the generated business insights to address the specific business needs compared to their same-industry competitors (Campher, 2022; Rehman et al., 2018). Besides, there is an inevitable positive connection between Extract, Transform and Load (ETL) processes and their significant impact on the performance of the parallel processed data warehouse (Bani et al., 2018). Therefore, a cloud data warehouse is one possible recommendation to accommodate the need for fast query performance (Bani et al., 2018; Rehman et al., 2018). On the other hand, star schema can be recommended to be used as the developmental foundation of an efficient data warehouse over snowflake schema in terms of query performance optimization based on the number of joins for foreign keys (Mohammed, 2014).

4.3 Data Integration and Quality

The third evaluation criteria of data warehouse concepts revolve around data integration and quality. The association between data quality and the subsequent integration of data from multiple sources into the data warehouse is important in the development of a holistic data warehouse (Mohammed, 2014). An efficient airline data warehouse is obliged to provide integrated, high-quality data to business stakeholders so that they can utilise these data to give themselves a competitive edge over other airlines by making the most accurate business decisions possible based on the generated business insights (Bani et al., 2018). During the integration of different types of data from multiple sources into the data warehouse through the ETL process, the technical users of the airline company need to ensure that these big data are of high quality on a timely and thorough basis, that is without data inconsistencies and incompleteness based on the quality standard defined based on specific subjects because once integrated and loaded into the data warehouse, any changes or updates cannot be made to the data due to its non-volatile nature, thus ensuring the simplicity and accuracy of data warehouse (Mohammed, 2014; Rehman et al., 2018; Revels & Nussbaumer, 2013). Therefore, addressing integrated historical data inconsistencies through timely data cleaning and preprocessing before loading them into the data warehouse is imperative for more accurate database query performance.

4.4 Data Security

The fourth evaluation criterion of data warehouse concepts is data security. The airline data warehouse is filled with highly private and sensitive data, especially from internal, passengers and competitors, in addition to the fact that it is linked to the Internet so practically the data inside the warehouse can be accessed by anybody including competitors and cross-departmental unauthorized internal staff regardless of legality (Arora & Gosain, 2020; Revels & Nussbaumer, 2013). Besides, these data are gathered from airline passengers through web mining without their consent before they are retrieved from the data warehouse to be utilized at the disposal of the airline business and the governmental agencies specialized in nationwide transportation (Revels & Nussbaumer, 2013). All these greatly impose data security threats to the passengers and the company itself, in addition to the fact that there is a possible vulnerability to data warehouse contamination, prompting the airline company to implement the appropriate data protection measures to embrace data confidentiality (Revels & Nussbaumer, 2013). Therefore, one way to evaluate the data security of data warehouses is the implementation of access and version control of data which provides accessibility of specific business areas to authorised internal staff belonging to the same field, in addition to facilitating monitoring of authorised timely changes made to the data before loading into the data warehouse as well as preparation for contingency measures like data backup (Arora & Gosain, 2020; Singh et al., 2019).

4.5 Flexibility and Support for Analytics

The final evaluation criterion for the data warehouse is flexibility and support for analytics. In terms of subject orientation, an airline data warehouse is capable of organizing and generating specific yet adaptive information that corresponds with the specific subject area within the airline business based on the dynamically evolving business needs, thus providing flexibility and support in analysing this information using a wide variety of output-efficient big data analytical tools and generating a more holistic view of business insights (Sinha, 2019). Besides, the non-volatile, integrated and time-variant nature of the data warehouse also indicates that it can flexibly accommodate different formats and types of past and present data within the database since the data arrives from different places and remains unmanipulable in the data warehouse over time, which can be an optimal support for further time-related analytics in a flexible manner (Sinha, 2019).

5.0 Conclusion

In conclusion, there is a need for airline companies to implement data warehouses which revolve around optimal scalability, performance, data integration, quality and security as well as analytical flexibility and support for better data-driven decision-making in the evolving aviation business landscape.

6.0 References

Antunes, A. L., Cardoso, E., & Barateiro, J. (2022). Incorporation of ontologies in data warehouse/business intelligence systems-a systematic literature review. *International Journal of Information Management Data Insights*, *2*(2), 100131. https://doi.org/10.1016/j.jjimei.2022.100131

Arora, A., & Gosain, A. (2020). Dynamic trust emergency role-based access control (DTE–RBAC). *International Journal of Computer Applications*, *175*(24), 20-24. https://doi.org/10.5120/ijca2020920773

Bani, F. C. D., Suharjito, Diana, & Girsang, A. S. (2018). Implementation of database massively parallel processing system to build scalability on process data warehouse. *Procedia Computer Science*, *135*, 68-79. https://doi.org/10.1016/j.procs.2018.08.151

Campher, S. E. (2022). Semantic metadata requirements for data warehousing from a dimensional modeling perspective. *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022)*, *1*, 129-136. https://doi.org/10.5220/0011018200003179

Mohammed, K. I. (2014). Data warehouse design and implementation based on quality requirements. *International Journal of Advances in Engineering & Technology*, *7*(3), 642-651. https://www.proquest.com/scholarly-journals/data-warehouse-design-implementation-based-on/docview/1550522116/se-2?accountid=46052

Rehman, K. U. U., Ahmad, U., & Mahmood, S. (2018). A comparative analysis of traditional and cloud data warehouse. *VAWKUM Transactions on Computer Sciences*, *6*(1), 34-40. https://doi.org/10.21015/vtcs.v15i1.487

Revels, M., & Nussbaumer, H. (2013). Data mining and data warehousing in the airline industry. *Academy of Business Research Journal*, *3*, 68-80. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2519737

Saddad, E., El-Bastawissy, A., Mokhtar, H. M., & Hazman, M. (2020). Lake data warehouse architecture for big data solutions. *International Journal of Advanced Computer Science and Applications*, *11*(8). https://doi.org/10.14569/IJACSA.2020.0110854

Singh, M. P., Sural, S., Vaidya, J., & Atluri, V. (2019). Managing attribute-based access control policies in a unified framework using data warehousing and in-memory database. *Computers & Security*, *86*, 183-205. https://doi.org/10.1016/j.cose.2019.06.001

Sinha, R. (2019). Analytical study of data warehouse. *International Journal of Management, IT & Engineering*, *8*(1), 105-115. http://doi.org/10.13140/RG.2.2.22600.65285