Final case Analysis

Shruti Daund

Business statistics 1

The final case analysis is based on aviation industry dataset itineraries sold in the first quarters of years 2009, 2017, 2018. The rates of flight tickets vary based on the month of the year for example the rates of tickets are high in December and august. The months selected are consistent for all the years mentioned.

The variables used in this dataset are:

ItnilD - unique ID for each transaction

Year – year when flight took place

Origin – starting destination airport

originStateName – state where airport is located

roundtrip- if its roundtrip then 1 or else 0

online – if booked online then 1 or else 0

FarePerMile- charges per mile

RPcarrier – the airline

Passengers

ItinFare

Distance

DistanceGroup

MilesFlown
Region
Division
Pop

Data cleaning

Conditions

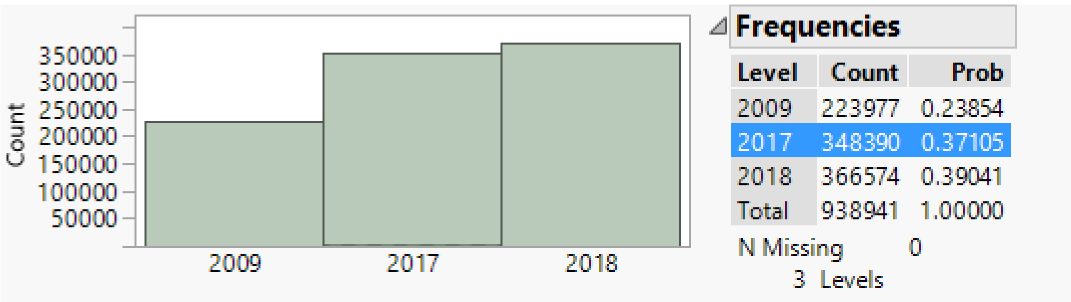
Continuous variables:

a) Itinfare **less than** 9,800
b) Passengers **less than** 100
c) Distance **less than** 10,000

Categorical Variables:

a) OriginStateNames: Remove “U.S. Pacific Trust Territories and Possessions”, “Puerto Rico” and “U.S. Virgin Islands”
b) RPCarrier: remove all carriers that have less than 5000 itineraries listed in the dataset

The data was cleaned by applying appropriate filters for 3 years combined. After getting the subset we get total number of rows as 938,941. We can observe variance in the overall tickets sold in these 3 years. 2009 shows a lower count since it is the year following the economic depression. 2017 shows growth in the sales of the tickets but the number of tickets being sold in 2018 as compared to 2017 is not impressing as there is only slight growth of 18,184.

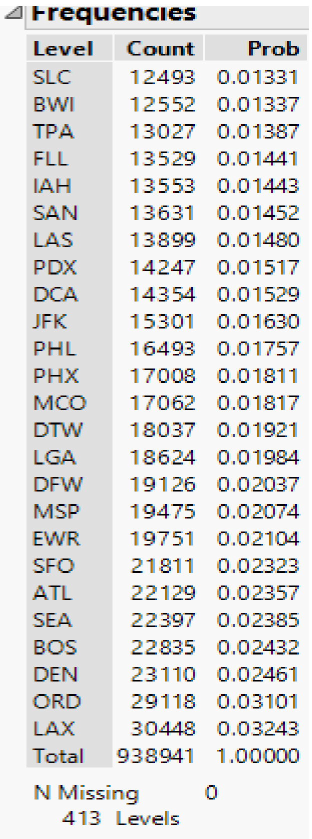


Frequencies

Level	Count	Prob
2009	223977	0.23854
2017	348390	0.37105
2018	366574	0.39041
Total	938941	1.00000

N Missing 0
3 Levels

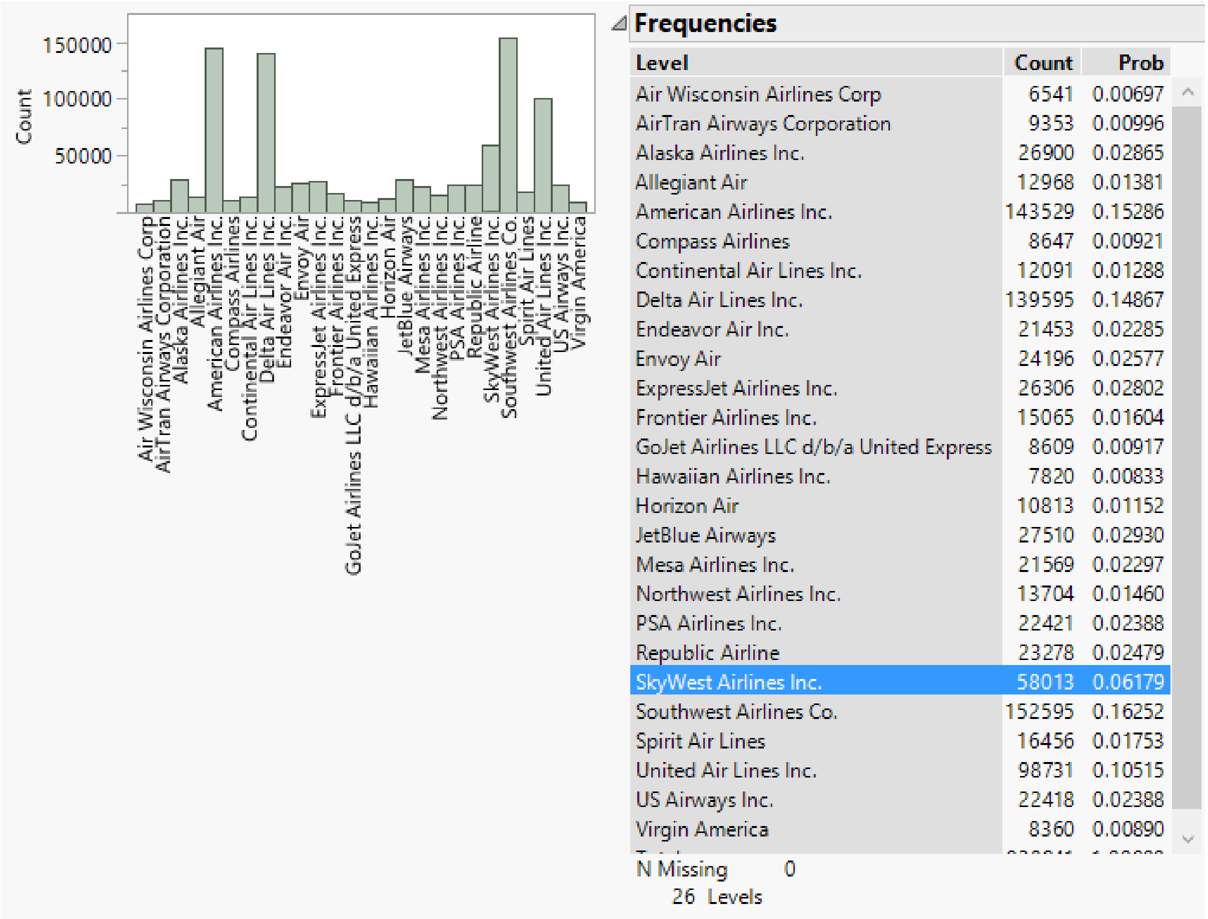
The number of flights taking off from LAX airport are 30448 which is highest than any other airport that is 30448. And the lowest are in AIA that is 1.



Frequencies

Level	Count	Prob
SLC	12493	0.01331
BWI	12552	0.01337
TPA	13027	0.01387
FLL	13529	0.01441
IAH	13553	0.01443
SAN	13631	0.01452
LAS	13899	0.01480
PDX	14247	0.01517
DCA	14354	0.01529
JFK	15301	0.01630
PHL	16493	0.01757
PHX	17008	0.01811
MCO	17062	0.01817
DTW	18037	0.01921
LGA	18624	0.01984
DFW	19126	0.02037
MSP	19475	0.02074
EWR	19751	0.02104
SFO	21811	0.02323
ATL	22129	0.02357
SEA	22397	0.02385
BOS	22835	0.02432
DEN	23110	0.02461
ORD	29118	0.03101
LAX	30448	0.03243
Total	938941	1.00000

N Missing 0
413 Levels



Frequencies

Level	Count	Prob
Air Wisconsin Airlines Corp	6541	0.00697
AirTran Airways Corporation	9353	0.00996
Alaska Airlines Inc.	26900	0.02865
Allegiant Air	12968	0.01381
American Airlines Inc.	143529	0.15286
Compass Airlines	8647	0.00921
Continental Air Lines Inc.	12091	0.01288
Delta Air Lines Inc.	139595	0.14867
Endeavor Air Inc.	21453	0.02285
Envoy Air	24196	0.02577
ExpressJet Airlines Inc.	26306	0.02802
Frontier Airlines Inc.	15065	0.01604
GoJet Airlines LLC d/b/a United Express	8609	0.00917
Hawaiian Airlines Inc.	7820	0.00833
Horizon Air	10813	0.01152
JetBlue Airways	27510	0.02930
Mesa Airlines Inc.	21569	0.02297
Northwest Airlines Inc.	13704	0.01460
PSA Airlines Inc.	22421	0.02388
Republic Airline	23278	0.02479
SkyWest Airlines Inc.	58013	0.06179
Southwest Airlines Co.	152595	0.16252
Spirit Air Lines	16456	0.01753
United Air Lines Inc.	98731	0.10515
US Airways Inc.	22418	0.02388
Virgin America	8360	0.00890

N Missing 0
26 Levels

As we can see from the image above the Airlines which most of the people preferred travelling with was southwest airlines Co with count of 15,2595. And the least preferred one was AirWisconsin airline Co with a count of 6541.

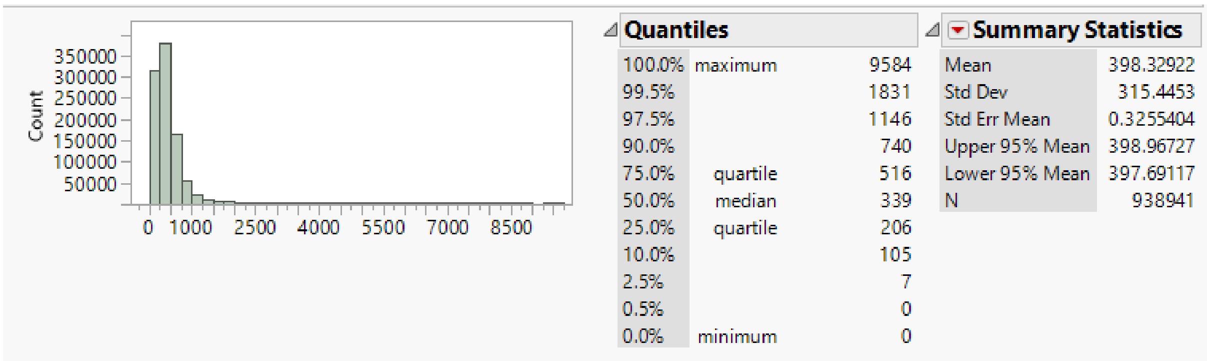
Data Analysis

Univariate Analysi

Describe shape and other descriptive statistics of Itinfare, Distance and Passengers variables

Itinfare

As we can see from the image the average fare for all the three years combined with a 95% confidence level is 398.32.



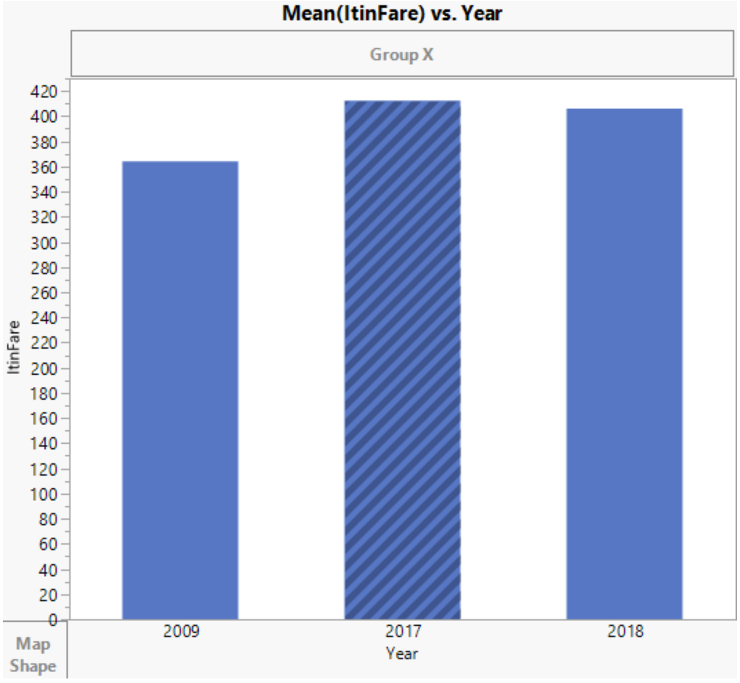
Quantiles

100.0%	maximum	9584
99.5%		1831
97.5%		1146
90.0%		740
75.0%	quartile	516
50.0%	median	339
25.0%	quartile	206
10.0%		105
2.5%		7
0.5%		0
0.0%	minimum	0

Summary Statistics

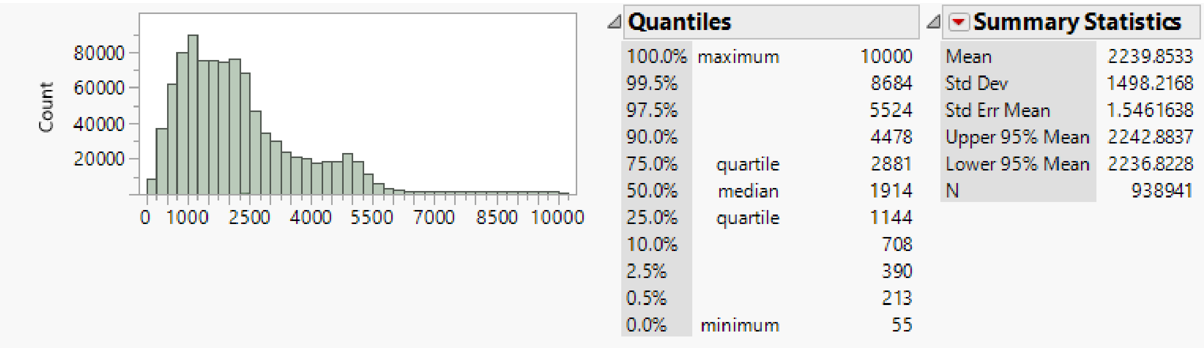
Mean	398.32922
Std Dev	315.4453
Std Err Mean	0.3255404
Upper 95% Mean	398.96727
Lower 95% Mean	397.69117
N	938941

As we can see the average fare has dropped in the year 2018 compared to the previous year 2017. The average fare in 2009, 2017 and 2018 was $364, $412 and $406, respectively



Distance

The average distance travelled in 3 years is 2239.8533 miles



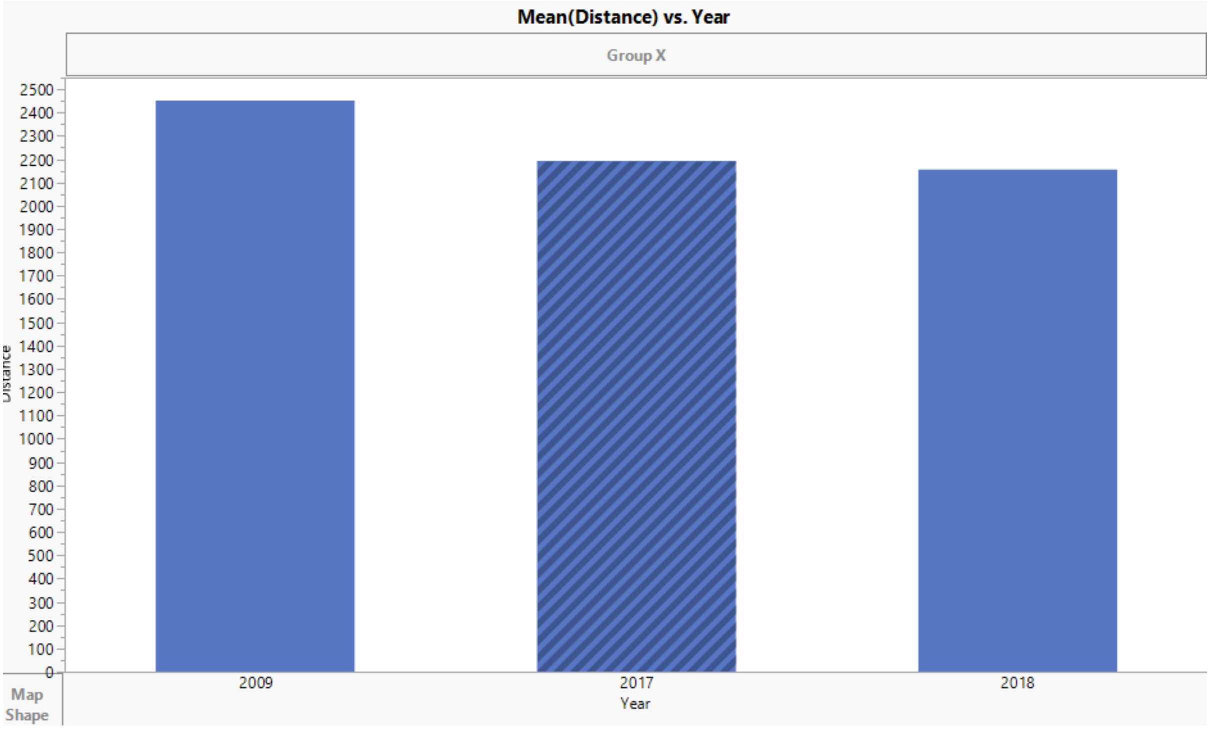
Quantiles

100.0%	maximum	10000
99.5%		8684
97.5%		5524
90.0%		4478
75.0%	quartile	2881
50.0%	median	1914
25.0%	quartile	1144
10.0%		708
2.5%		390
0.5%		213
0.0%	minimum	55

Summary Statistics

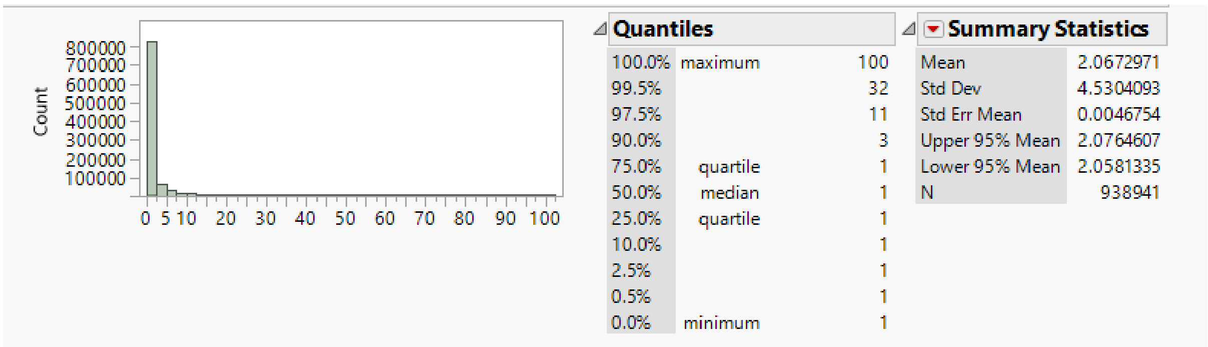
Mean	2239.8533
Std Dev	1498.2168
Std Err Mean	1.5461638
Upper 95% Mean	2242.8837
Lower 95% Mean	2236.8228
N	938941

We can see from the graph that there is a decrease in the distance travelled with each year passing. Comparing the three years separately, the average distance travelled in 2009 was 2451 miles while it was 2192 miles in 2017 and 2156 miles in 2018



Passengers

The average number of passengers travelling for 3 years combined is 2.0672971 per itinerary. 90% of the times no of passenger travelling was 3.



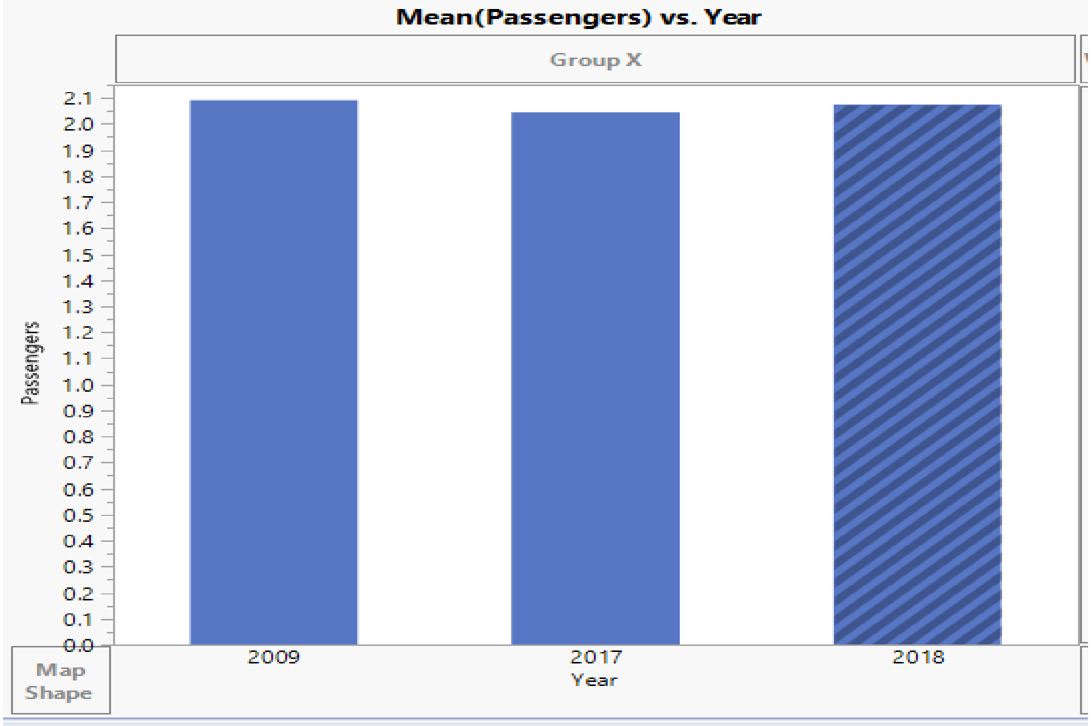
Quantiles

100.0%	maximum	100
99.5%		32
97.5%		11
90.0%		3
75.0%	quartile	1
50.0%	median	1
25.0%	quartile	1
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	1

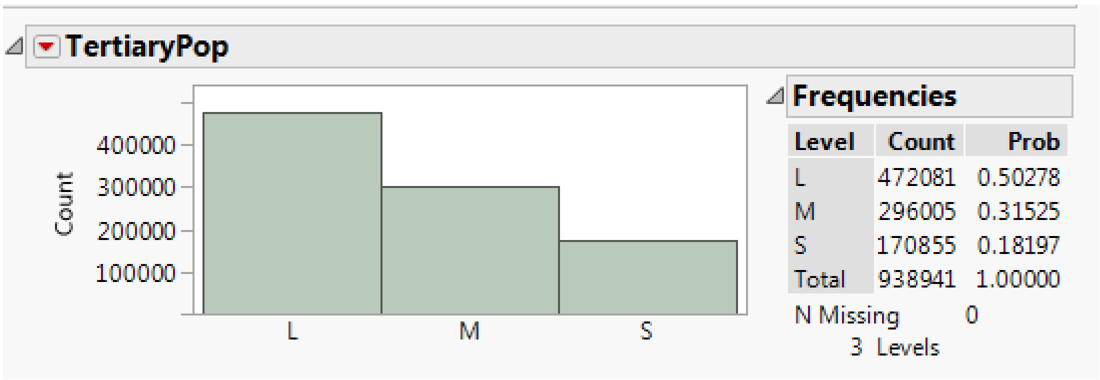
Summary Statistics

Mean	2.0672971
Std Dev	4.5304093
Std Err Mean	0.0046754
Upper 95% Mean	2.0764607
Lower 95% Mean	2.0581335
N	938941

There is only slight difference in the number of passengers travelling in these 3 years. It has been consistent at 2.



2.Create a new tertiary (three categories) flag variable using population "Pop" variable for states that have populations more than 10 million, 5-10 million and those less than that. Call this variable TertiaryPop (stands for Tertiary population).



Frequencies

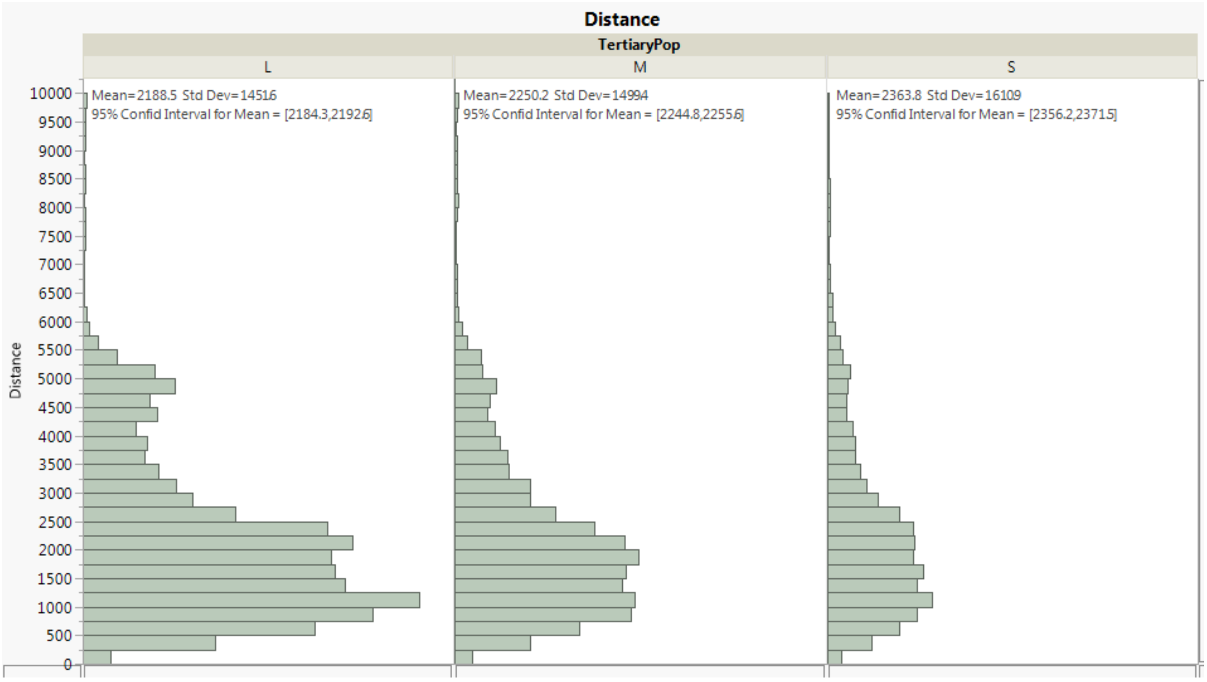
Level	Count	Prob
L	472081	0.50278
M	296005	0.31525
S	170855	0.18197
Total	938941	1.00000

N Missing 0

3 Levels

A new column was created by sorting states that have population more than 10 million, population between 5-10 million and less than that.

3. Run descriptive analysis including confidence intervals on Distance, Itinfare and FareperMile and comment if there are differences in the above by the size of the state



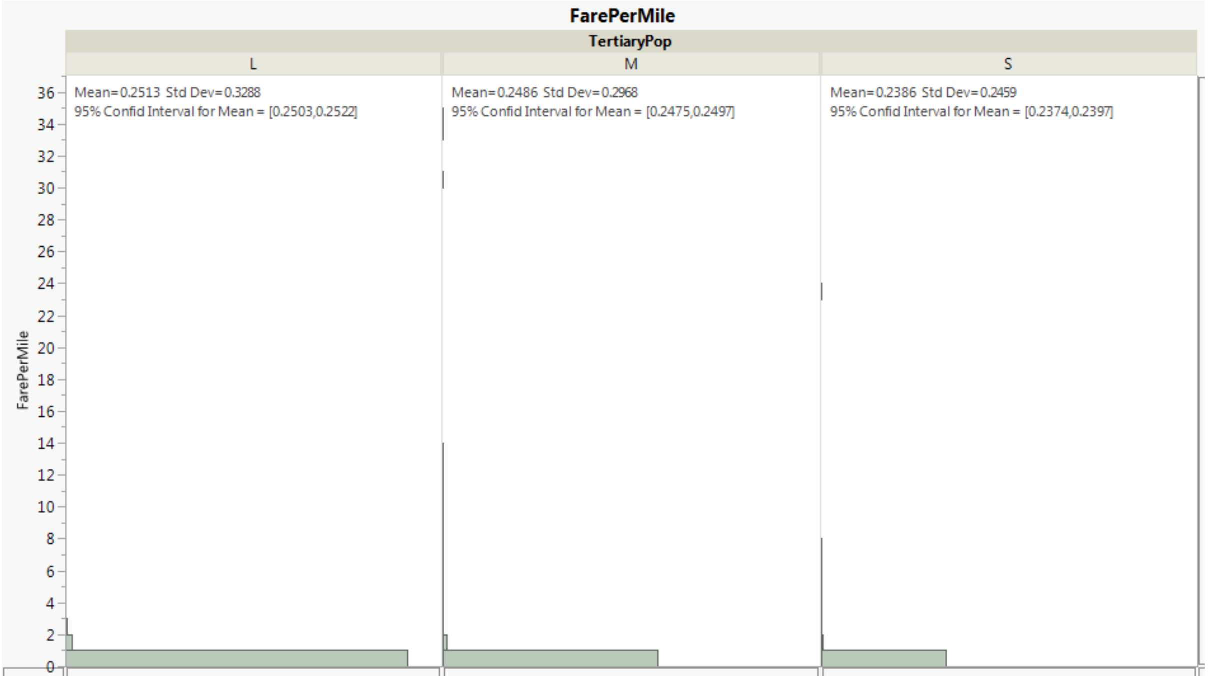
Distance-

From the above graph we can state that the average distance travelled for large states is 2188.5 miles 95% CI at [2184, 2192]), medium states is 2250.2 miles (CI- [2245, 2256]) and small states is 2363.8 miles Confidence Interval- (2356, 2372). Thus with increase in the area of the state, the average distance travelled has decreased.

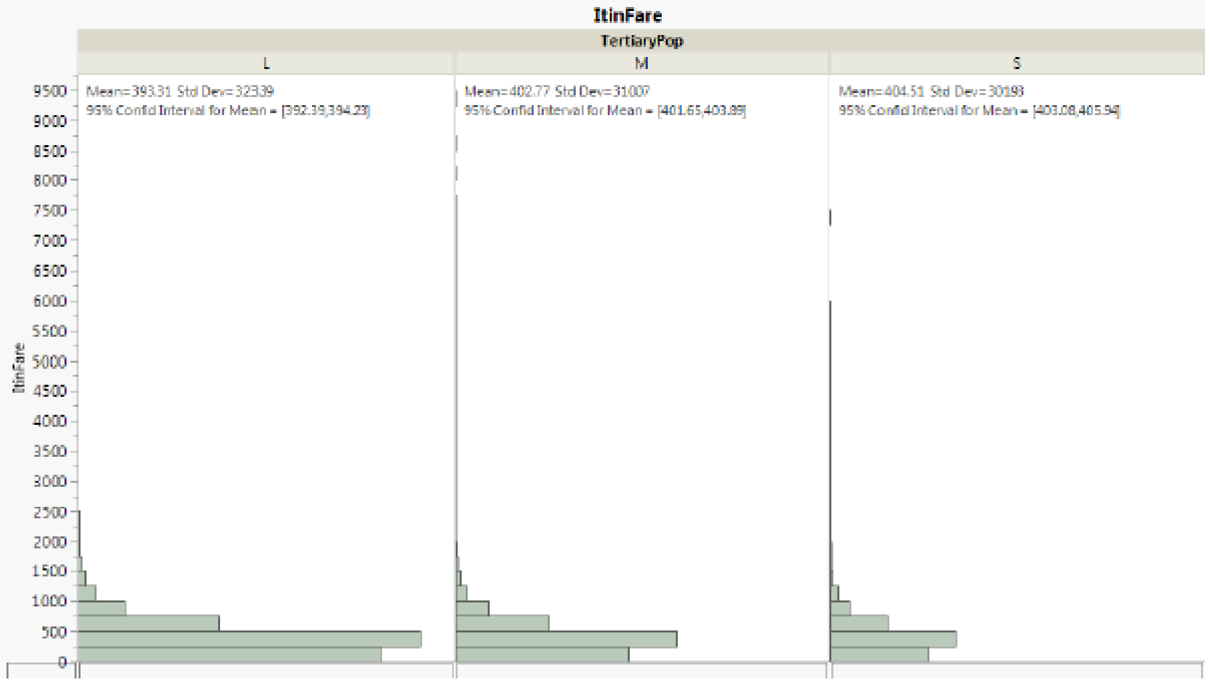
FarePerMile

The fare per mile has decreased with the size of the state but the difference is very small as we can observe from the graph.

CI for large states- [0.2503, 0.2522]
CI for medium states- [0.2475, 0.2497]
CI for small states- [0.2374, 0.2397]



ItinFare



According to the graph as the size of the state increases the fare decreses.

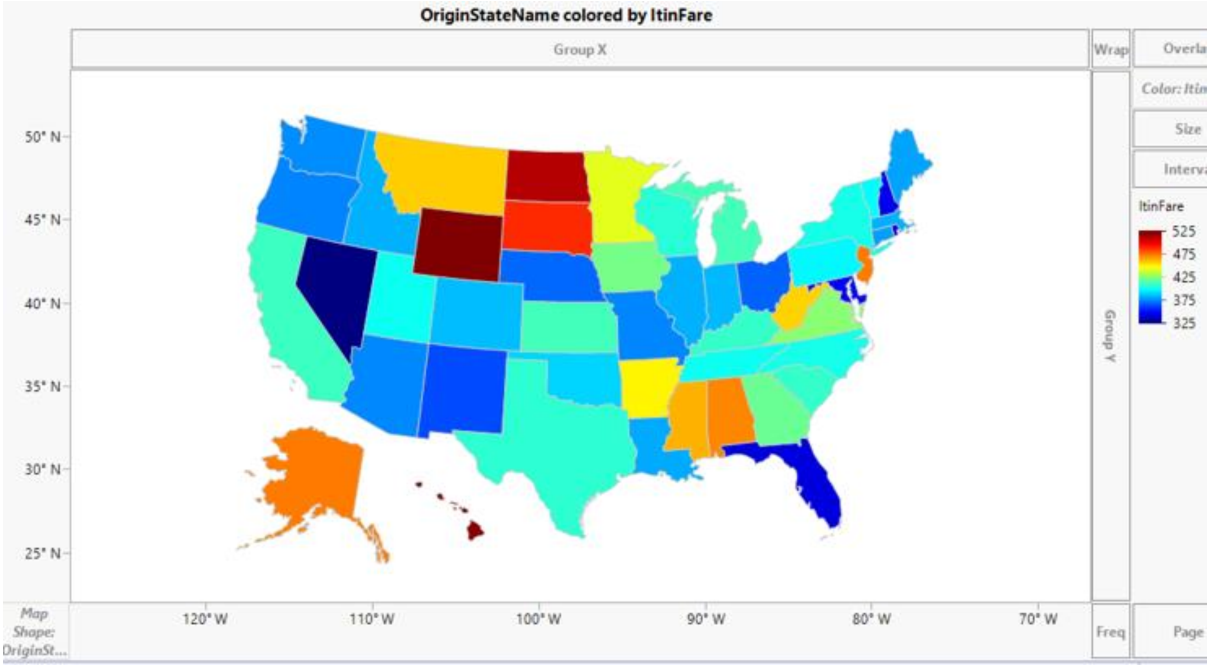
Tickets from large states cost an average of $393 with a 95% CI of [392, 394]. Medium sized states had their tickets priced at an average of $403 with CI of [402, 404] and small states averaged at $405 with CI of [404, 406].

4. It is of interest for consumers to know if there are differences in airline ticket fares by the origin of the flight. In other words is there a difference in ticket fares to fly from different states? Run a comprehensive analysis of comparing average ticket price and average ticket price / mile by state and *visualize* the differences using mapping facility in JMP and comment on the trends.

a. For all three years combined
b. Separately for three years

itinFare

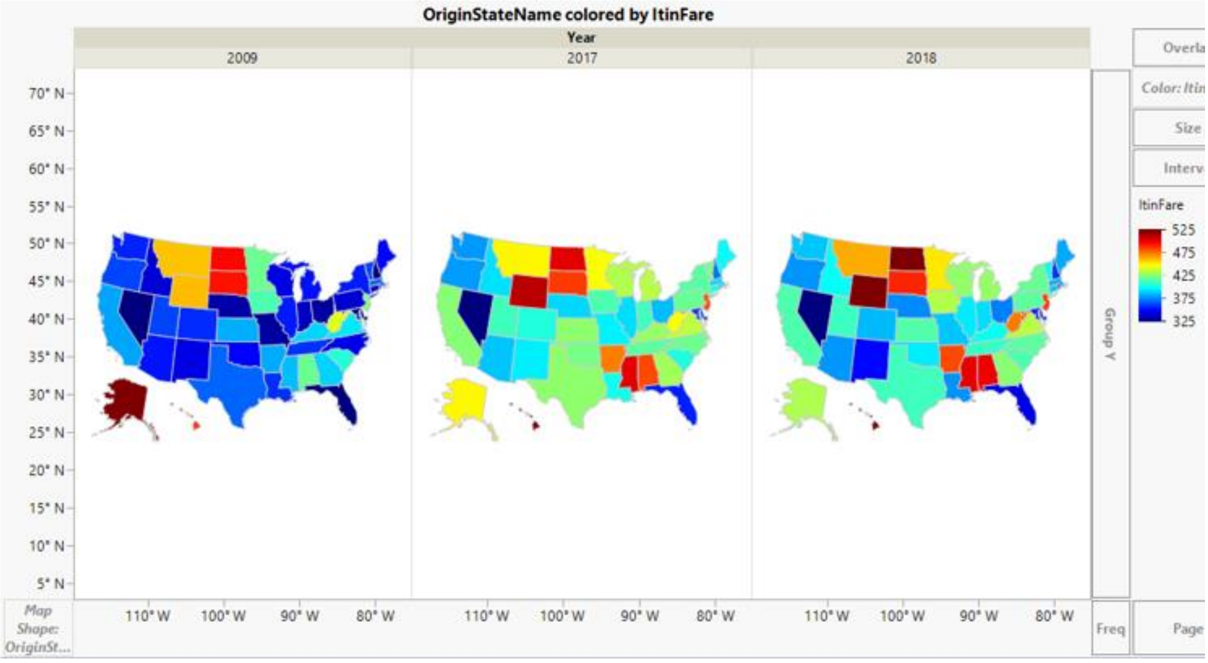
1)



Wyoming has the highest ticket fare of 528$ followed by North Dakota. Nevada has the cheapest fares of 321$ followed by florida.

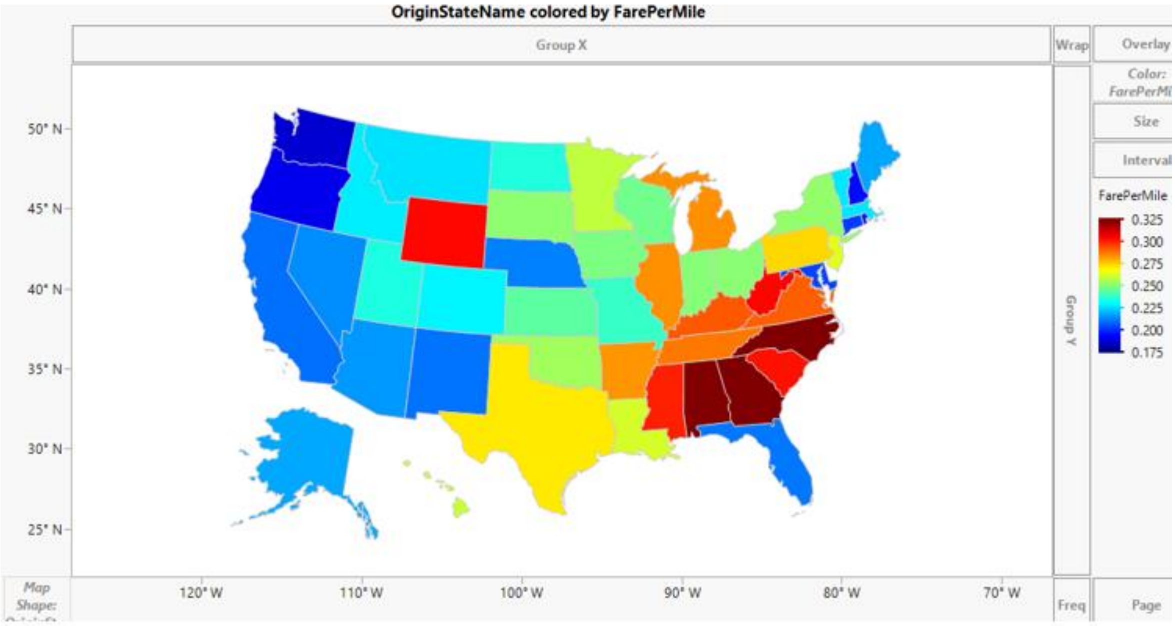
2) three years combined

As seen in the graph in 2009, Alaska was the most expensive with a fare of $573 average while Florida was the cheapest at $319. In 2017, Hawaii was on top for the highest ticket rates at $535 while Nevada became the cheapest at $324. In 2018, Wyoming was the most expensive source of flight with average prices at $587 as Nevada remained as the cheapest at $318.



FarePerMile

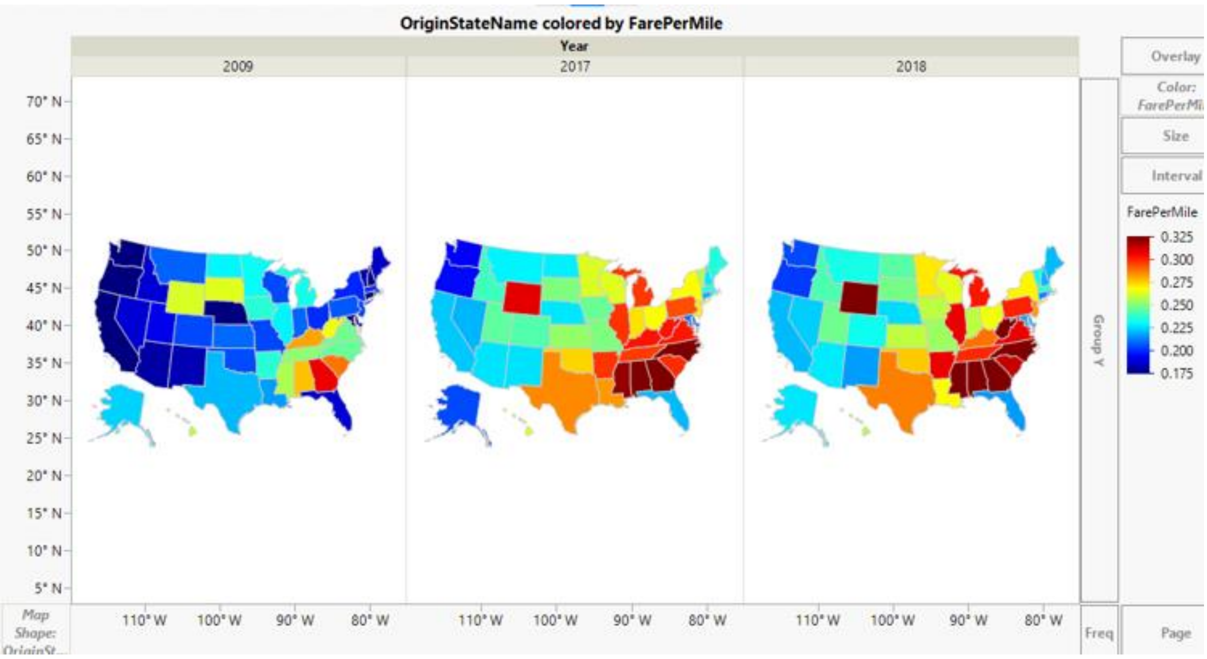
1)



North Carolina had the highest fare per mile at $0.3296 and Washington State had the lowest at $0.1870.

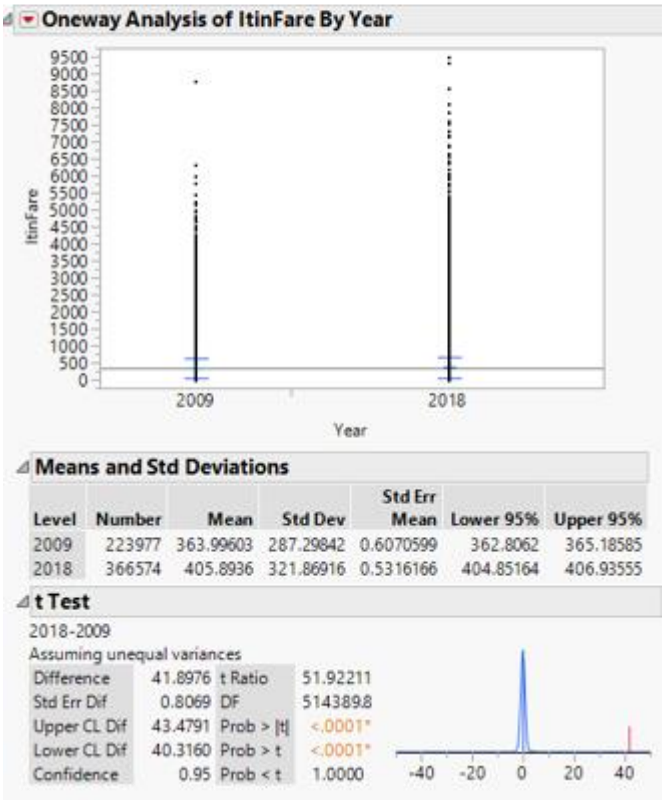
2) 3 years combined

As seen in the graph in 2009, New Hampshire had the lowest fare per mile at $0.1448 and Georgia had the highest at $0.3090. In 2017, Washington State was the cheapest at $0.1932 while North Carolina was highest with $0.3463. In 2018, California was the lowest $0.2024 and North Carolina remained at the top at $0.3583.



Bivariate Analysis

a) **Have prices changed over time** Is there a significant difference in airfare for years 2018 and 2009? If yes why, if not why not?



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
2009	223977	363.99603	287.29842	0.6070599	362.8062	365.18585
2018	366574	405.8936	321.86916	0.5316166	404.85164	406.93555

t Test

2018-2009

Assuming unequal variances

Difference	41.8976	t Ratio	51.92211
Std Err Dif	0.8069	DF	514389.8
Upper CL Dif	43.4791	Prob > |t|	<.0001*
Lower CL Dif	40.3160	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

Condition

i. $H_0: \mu_{2018} = \mu_{2009}$

$H_1: \mu_{2018} \neq \mu_{2009}$

ii. $H_0: \mu_{2018} \geq \mu_{2009}$

$H_1: \mu_{2018} < \mu_{2009}$

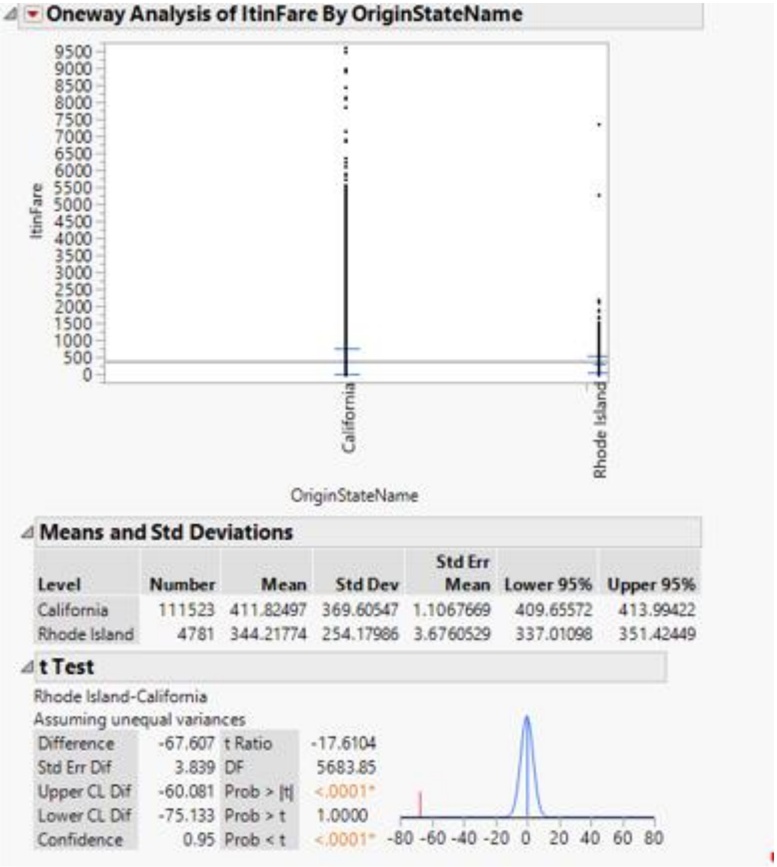
iii. $H_0: \mu_{2018} \leq \mu_{2009}$

$H_1: \mu_{2018} > \mu_{2009}$

The p value for Prob > |t| is .0001, which is negligible so the first condition that both the averages are equal can be rejected. Similarly the 2nd condition that average of 2018 is greater than the average of 2009 can also be rejected. However, the p-value of 1.0000 is very high for Prob < t, that means the null hypothesis that average of 2018 is lesser than that of 2009, has to be considered.

We can state that the average itinerary fare for 2009 was less than 2018. Prices have changed overtime, prices have increased over the years.

b) **Is it more expensive to fly out of smaller states** Test the theory on average ticket prices and price/mile for the largest (California) and the smallest (Rhode Island) states in the nation.



Oneway Analysis of ItinFare By OriginStateName

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
California	111523	411.82497	369.60547	1.1067669	409.65572	413.99422
Rhode Island	4781	344.21774	254.17986	3.6760529	337.01098	351.42449

t Test

Rhode Island-California

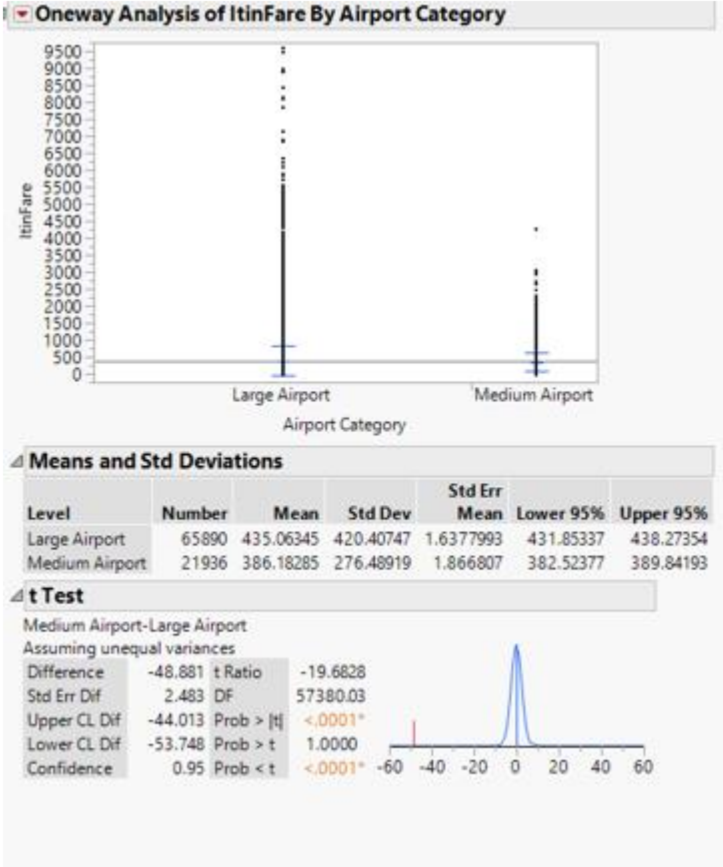
Assuming unequal variances

Difference	-67.607	t Ratio	-17.6104
Std Err Dif	3.839	DF	5683.85
Upper CL Dif	-60.081	Prob > |t|	<.0001*
Lower CL Dif	-75.133	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

The p-value for Prob > t is 1.0000 the p-value for the hypothesis that the airfare from Rhode Island is less than or equal to that of California has a very high value and cannot be rejected. The null hypotheses that the values are equal and that airfare from Rhode Island is greater than or equal to that of California have both p-values of .0001 which is very less. Here it cannot be proved that it is expensive to fly from smaller states.

Some airports within the same state are more expensive to fly out of- State of California has ~ 30 airports with multiple airports present in larger cities. LAX, SAN, and SFO are three of the largest airports and SNA, SMF and OAK are three medium size airports. Combine the largest and medium size airports in California into two categories and perform a test of hypothesis on the average airfare for the two groups.

.



Oneway Analysis of ItinFare By Airport Category

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Large Airport	65890	435.06345	420.40747	1.6377993	431.85337	438.27354
Medium Airport	21936	386.18285	276.48919	1.866807	382.52377	389.84193

t Test

Medium Airport-Large Airport

Assuming unequal variances

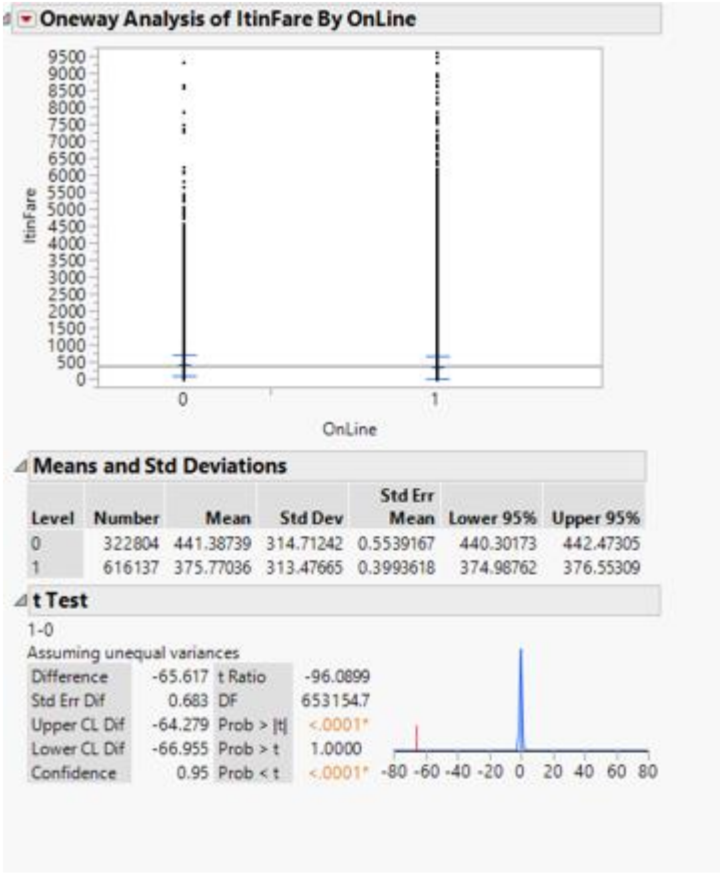
Difference	-48.881	t Ratio	-19.6828
Std Err Dif	2.483	DF	57380.03
Upper CL Dif	-44.013	Prob > |t|	<.0001*
Lower CL Dif	-53.748	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

The null hypothesis that airfares from medium airports are less than or equal to airfares from large airports has the high p-value of 1.0000 and cannot be rejected. The null hypotheses that the values are equal and that the airfares from medium airports are greater than or equal to airfares from larger airports can be ignored on the basis of p-value of .0001 in favour of the hypotheses that they are unequal and that airfares from medium airports are cheaper than that of large airports, **so it proves that large airports from California are expensive to fly out compared to medium sized airport.**

4) **Online purchases are cheaper than otherwise** Run tests of hypothesis to check if there are significant differences between airfares and miles flown for itineraries that were either purchased online or not.

AirFare

The null hypothesis that online tickets are cheaper than or equal to tickets purchased offline has value of 1.0000 cannot be rejected. However, hypotheses that online tickets are more expensive or equal to offline tickets can be ignored because of p-value of .0001. This means that the alternate hypothesis that **online tickets are cheaper than offline tickets** is true.



Oneway Analysis of ItinFare By OnLine

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
0	322804	441.38739	314.71242	0.5539167	440.30173	442.47305
1	616137	375.77036	313.47665	0.3993618	374.98762	376.55309

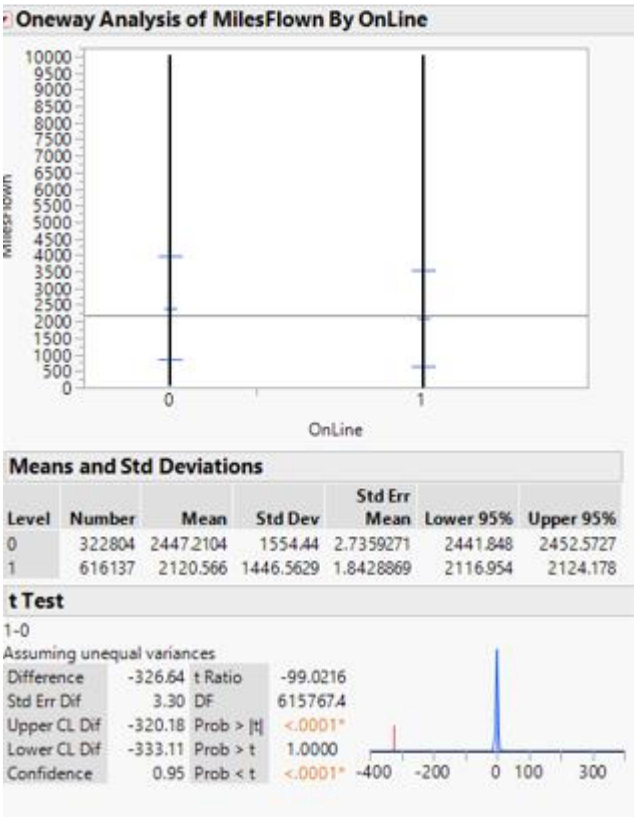
t Test

1-0

Assuming unequal variances

Difference	-65.617	t Ratio	-96.0899
Std Err Dif	0.683	DF	653154.7
Upper CL Dif	-64.279	Prob > |t|	<.0001*
Lower CL Dif	-66.955	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

Miles flown



Oneway Analysis of MilesFlown By OnLine

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
0	322804	2447.2104	1554.44	2.7359271	2441.848	2452.5727
1	616137	2120.566	1446.5629	1.8428869	2116.954	2124.178

t Test

1-0

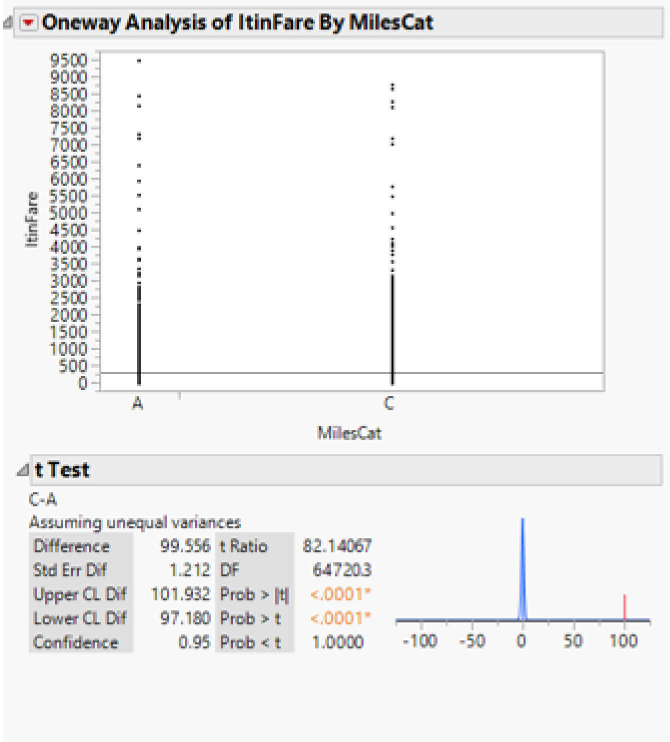
Assuming unequal variances

Difference	-326.64	t Ratio	-99.0216
Std Err Dif	3.30	DF	615767.4
Upper CL Dif	-320.18	Prob > |t|	<.0001*
Lower CL Dif	-333.11	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

The p-value for the null hypothesis that online tickets have higher or equal miles as that of tickets purchased offline is .0001. So this null hypothesis can be rejected of the hypothesis that **online tickets have lower miles flown compared to offline tickets**.

c) **The longer the flight, the pricier it is** Test this theory by creating a four category variable from MilesFLown variable: “< 500 miles”, “500 – 1200 miles”, “1200 – 2000 miles”, “2000 - 3000 miles”, “3000+ miles”. Run the following tests of hypothesis:

   a. Is there a difference in average fare cost between flights that flew “<500 miles” and those that flew “1200-2000 miles”?



Oneway Analysis of ItinFare By MilesCat

t Test

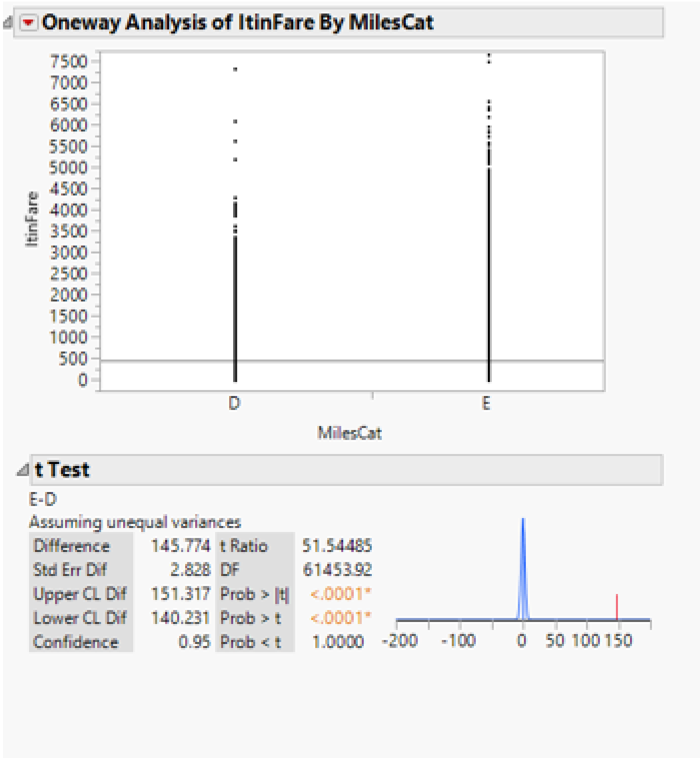
C-A

Assuming unequal variances

Difference	99.556	t Ratio	82.14067
Std Err Dif	1.212	DF	647203
Upper CL Dif	101.932	Prob > |t|	<.0001*
Lower CL Dif	97.180	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

The null hypothesis that average airfare for flights that flew less than 500 miles is greater than or equal to those that flew between 1200 and 2000 miles is .0001 and can be rejected. And the hypothesis that **that the average airfare for flights that flew between less than 500 miles is less than those that flew between 1200 and 2000 miles has a high p-value and has to be considered.**

   b. Test the hypothesis that there is a difference in average fare cost between flights that flew “2000 – 3000 miles” and “3000+ miles” in the Northeast region.



Oneway Analysis of ItinFare By MilesCat

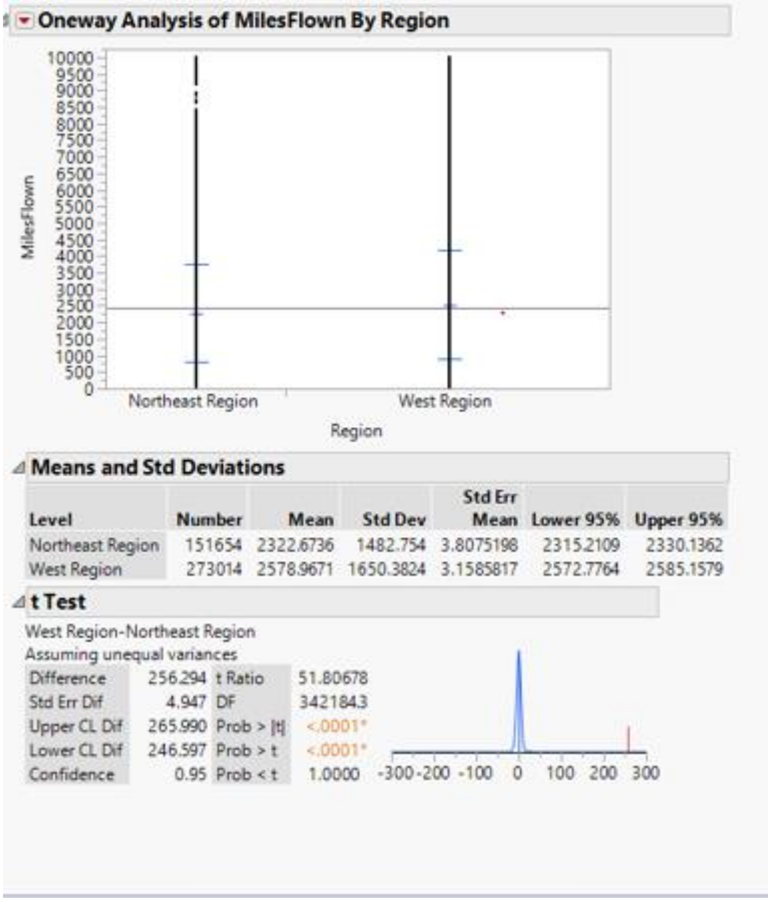
t Test

```
E-D
Assuming unequal variances
Difference     145.774  t Ratio     51.54485
Std Err Dif      2.828  DF          61453.92
Upper CL Dif   151.317  Prob > |t|   <.0001*
Lower CL Dif   140.231  Prob > t     <.0001*
Confidence        0.95  Prob < t      1.0000
```

From the above image we can see that the p-value for the null hypothesis that average fare cost between flights that flew between 2000 and 3000 miles is greater than or equal to that of 3000+ miles in the Northeast region is .0001. Hence, this hypothesis can be rejected and the hypothesis **that average fare cost between flights that flew "2000 – 3000 miles" is less than "3000+ miles" in the Northeast region has to be considered.**

3. Run a test of hypothesis to see if there is a significant difference in the miles flown from airports in the Northeast region compared to the Western region.

The t test shows that the p-value for the hypothesis that the miles flown from airports in the Northeast region is higher or equal compared to the Western region is .0001. Hence, it proves that **miles flown from airports in the Northeast region is less compared to the Western region.**



Oneway Analysis of MilesFlown By Region

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Northeast Region	151654	2322.6736	1482.754	3.8075198	2315.2109	2330.1362
West Region	273014	2578.9671	1650.3824	3.1585817	2572.7764	2585.1579

t Test

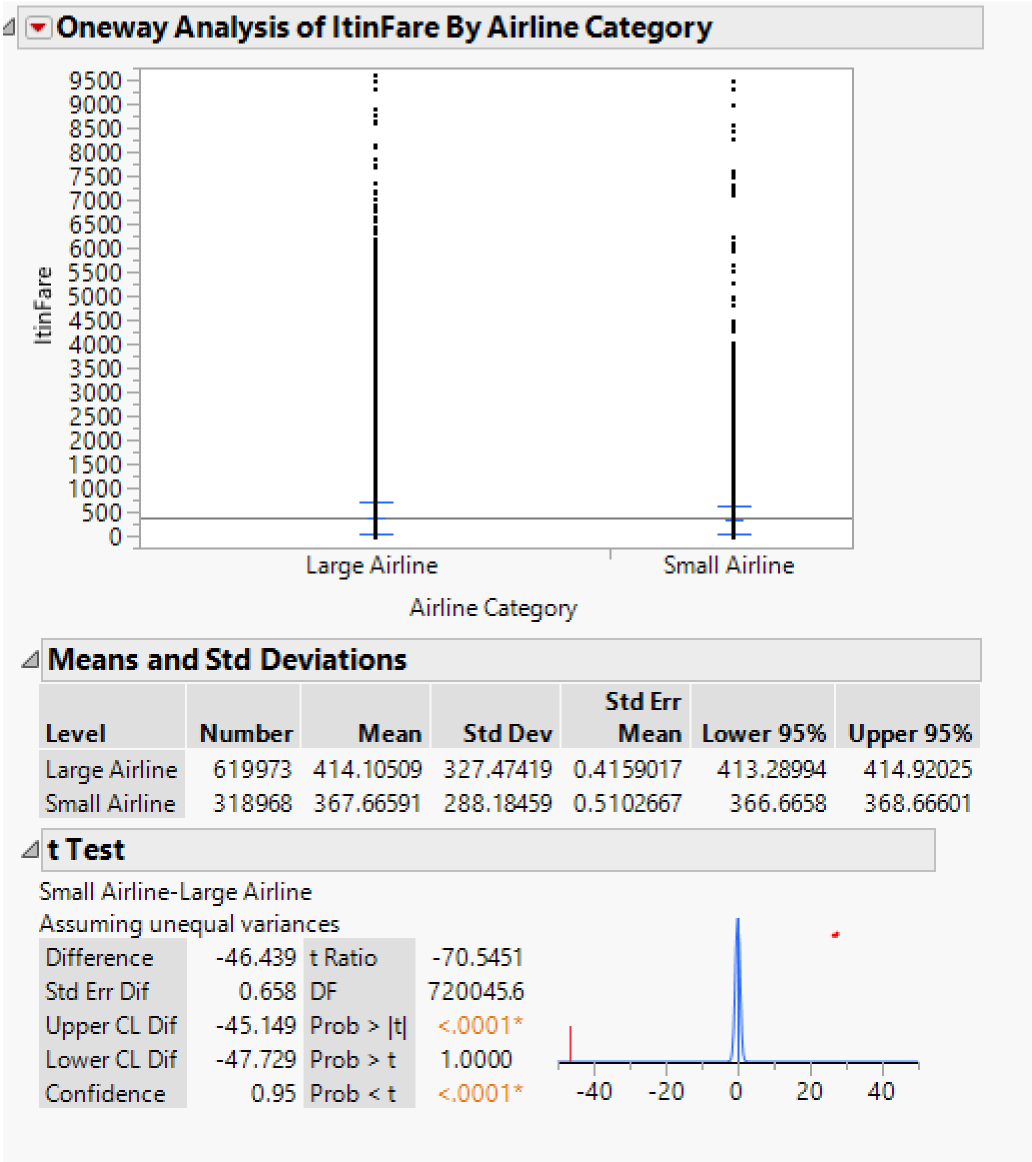
West Region-Northeast Region

Assuming unequal variances

Difference	256.294	t Ratio	51.80678
Std Err Dif	4.947	DF	342184.3
Upper CL Dif	265.990	Prob > |t|	<.0001*
Lower CL Dif	246.597	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

d) **Economy of scales in the airline industry** Southwest, American, Delta, United, Skywest and JetBlue are six of the largest domestic airline companies. The last question of the case relates to studying the phenomenon of “economy of scales”, i.e. are larger airlines on average able to offer more affordable prices for tickets compared to smaller airlines. Create two groups: “Large airlines” and “Small airlines”. Large Airlines consists of the six companies mentioned above and the Medium/small group consists of all other airlines that have less than 10k rows of data in the filtered dataset.

Run a test of hypothesis testing if there is a significant difference in the average airfare price for the two groups



Oneway Analysis of ItinFare By Airline Category

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Large Airline	619973	414.10509	327.47419	0.4159017	413.28994	414.92025
Small Airline	318968	367.66591	288.18459	0.5102667	366.6658	368.66601

t Test

Small Airline-Large Airline

Assuming unequal variances

Difference	-46.439	t Ratio	-70.5451
Std Err Dif	0.658	DF	720045.6
Upper CL Dif	-45.149	Prob > |t|	<.0001*
Lower CL Dif	-47.729	Prob > t	1.0000
Confidence	0.95	Prob < t	<.0001*

The p-value for the null hypothesis that airfare of small airlines is greater than or equal to that of large airlines is .0001. Hence, the null hypothesis can be rejected. And the alternate hypothesis that **airfares of small airlines is lesser than airfares of larger airlines is to be considered.**