# FINALS CASE STUDY
# APPLIED BUSINESS STATS I

Instructor- Prof. Prashant Mittal
Submitted by- Deepa Vincent Thomas

## DATASET

A large sample (almost 1 million rows) of domestic itineraries sold in the first quarters (Jan-March) of years 2009, 2017 and 2018 in the aviation industry.

## VARIABLES-

- ItinID
- Year
- Origin
- OriginStateName
- RoundTrip
- OnLine
- FarePerMile
- RPCarrier
- Passengers
- ItinFare
- Distance
- DistanceGroup
- MilesFlown
- Region
- Division
- Pop

## DATA CLEANING

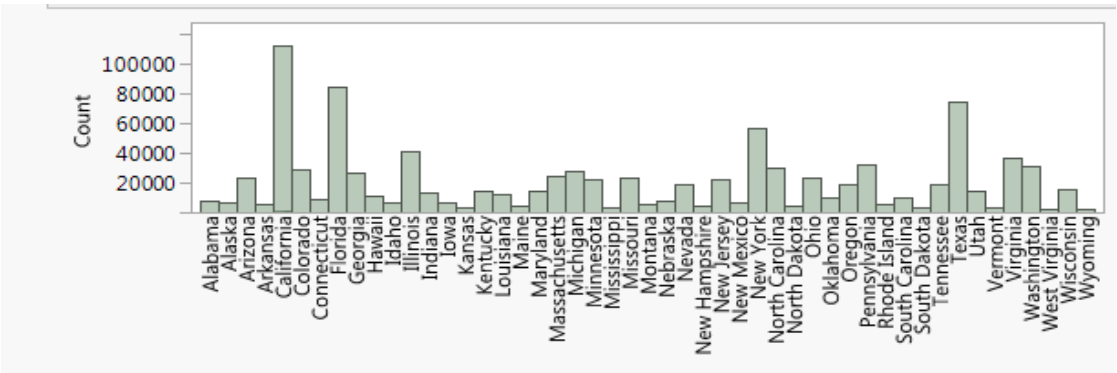
We were asked to clean the dataset based on the following conditions-

- Itinfare < 9,800
- Passengers < 100
- Distance < 10,000
- OriginStateNames- Remove "U.S. Pacific Trust Territories and Possessions", "Puerto Rico" and "U.S. Virgin Islands"
- RPCarrier- Remove all carriers that have less than 5000 itineraries listed in the dataset.

Once the filters were added a new subset of 938,941 rows was created out of the 971,206 rows in the given sample. In this dataset, the tickets were not equally distributed among the 3 years- 2009 had 24% of the total distribution while 2017 and 2018 were closer at 37 and 39% respectively. Since it was the first quarter after the economic depression in 2008, this result is not surprising.

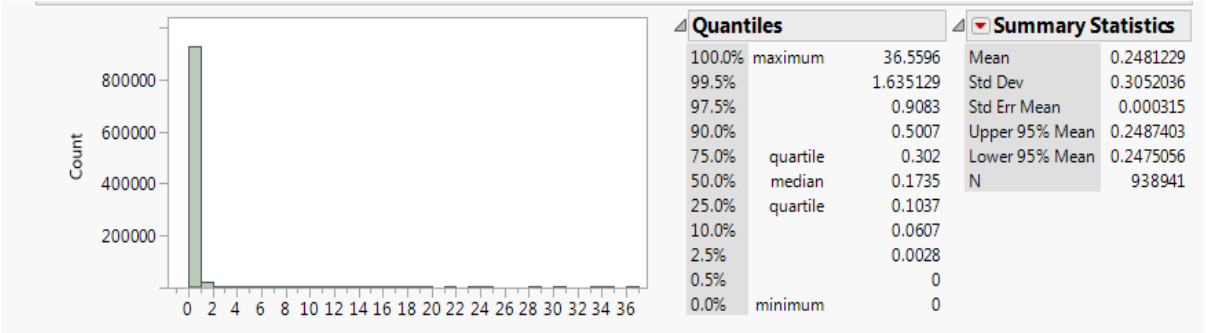
The following observations are for the 3 years combined, unless mentioned otherwise-

California had the biggest traffic in all 3 years combined with over 100,000 passengers and West Virginia had the lowest with just 833 passengers.

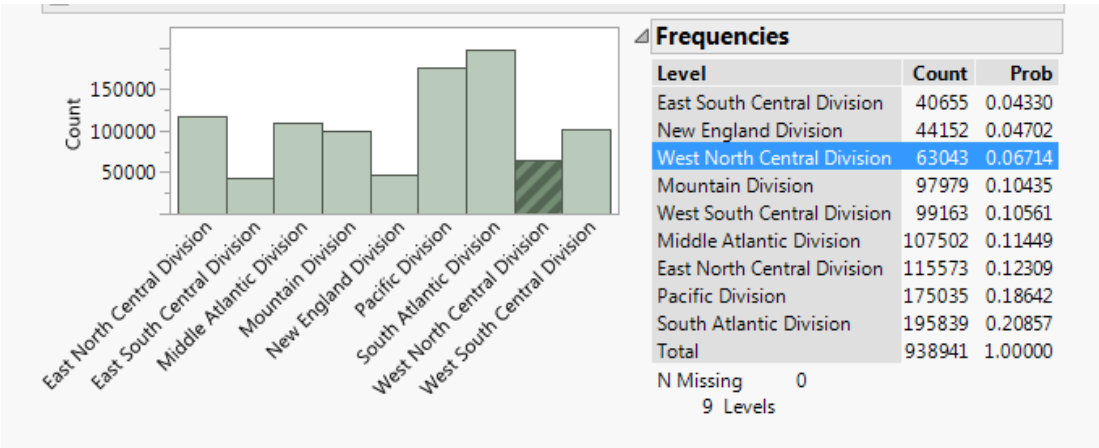



65% of the tickets were purchased online.

The average fare per mile was $0.25 with a maximum of $36.



Almost 35% of the flights took place in the south region followed by the west with 29%, Midwest 19% and northeast 16%. Breaking it down further, the most flights were in the South Atlantic division and the least was in the East South Central region.




**Frequencies**

| Level | Count | Prob |
|---|---|---|
| East South Central Division | 40655 | 0.04330 |
| New England Division | 44152 | 0.04702 |
| West North Central Division | 63043 | 0.06714 |
| Mountain Division | 97979 | 0.10435 |
| West South Central Division | 99163 | 0.10561 |
| Middle Atlantic Division | 107502 | 0.11449 |
| East North Central Division | 115573 | 0.12309 |
| Pacific Division | 175035 | 0.18642 |
| South Atlantic Division | 195839 | 0.20857 |
| Total | 938941 | 1.00000 |

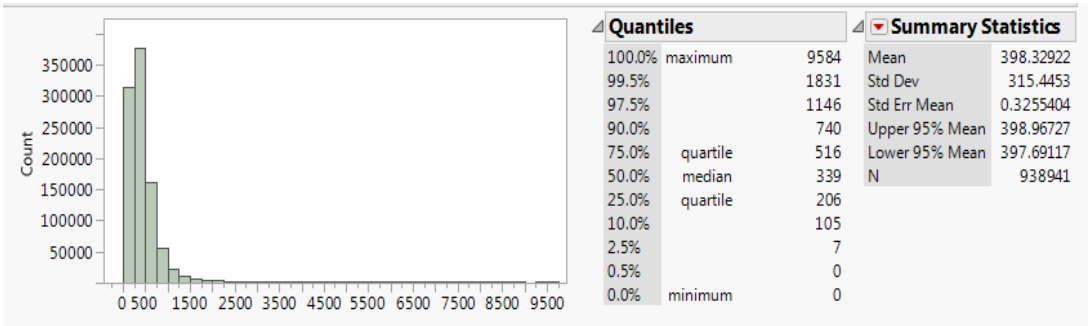
N Missing 0
9 Levels

## DATA ANALYSIS

Univariate Analysis

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

   ItinFare-
   The average fare per itinerary all the three years combined was $398.39 with a 95% Confidence Interval at (397.69, 398.97). The graph was right-skewed.



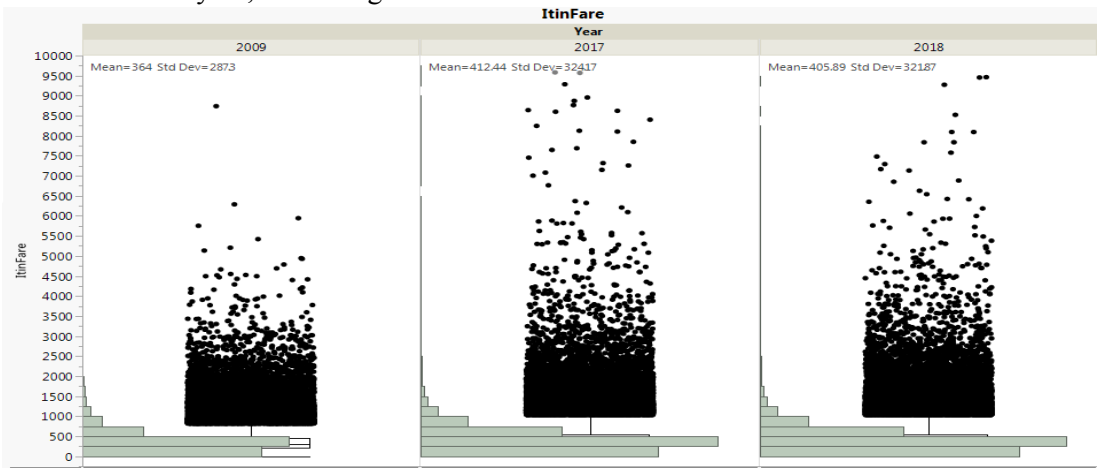

**Quantiles**

| | | |
|---|---|---|
| 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**

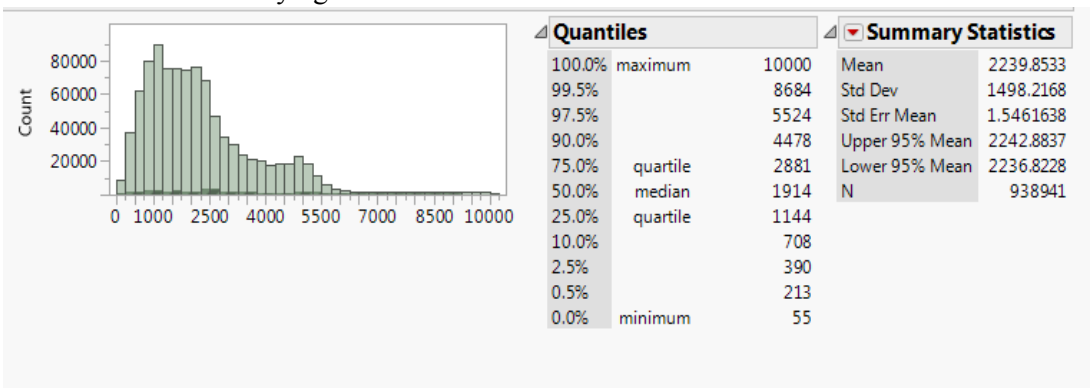
| | |
|---|---|
| Mean | 398.32922 |
| Std Dev | 315.4453 |
| Std Err Mean | 0.3255404 |
| Upper 95% Mean | 398.96727 |
| Lower 95% Mean | 397.69117 |
| N | 938941 |

The average fare in 2009, 2017 and 2018 was $364, $412 and $406, respectively. Rather than an increase with year, the average fare decreased from 2017 to 2018.



Distance-

The mean distance travelled 3 years combined was 2240 miles. The graph was right-skewed with 90% of the values lying between 55 and 4478 miles.



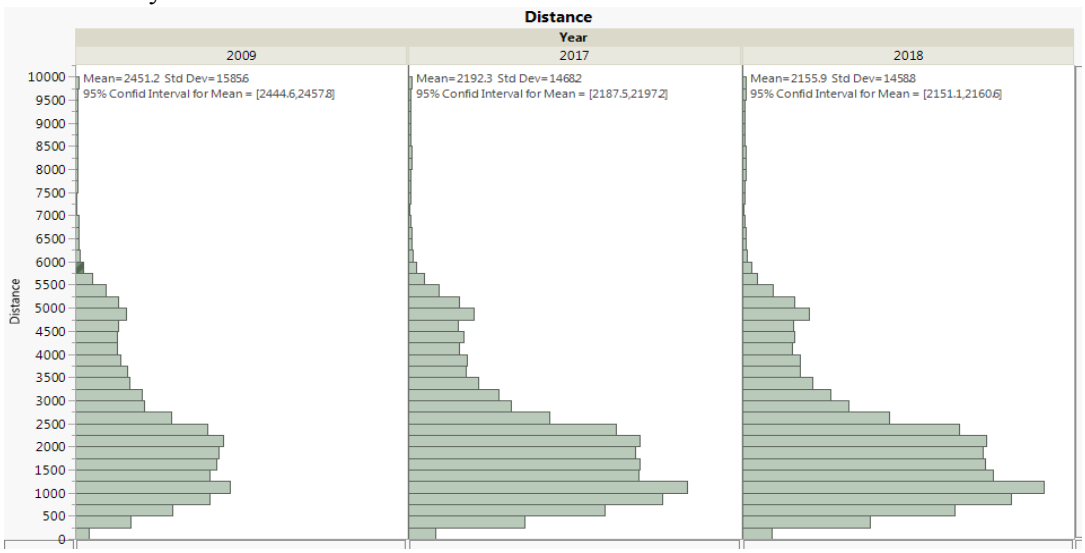

**Quantiles**

| | | |
|---|---|---|
| 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 |

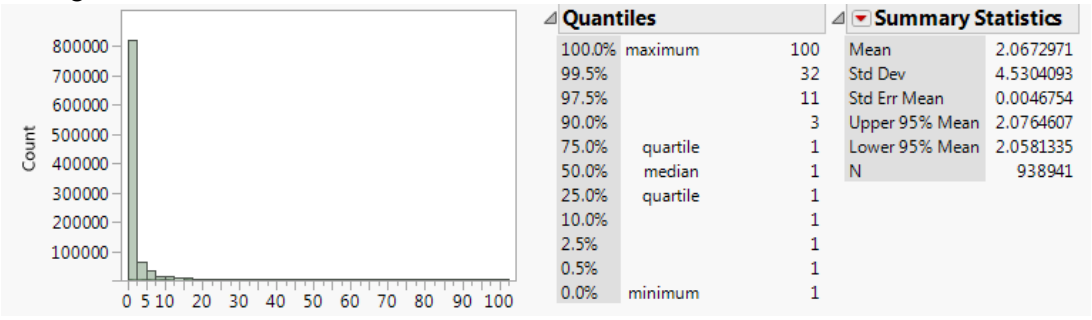
**Summary Statistics**

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

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. This shows a decrease in the variable with year.




Passengers



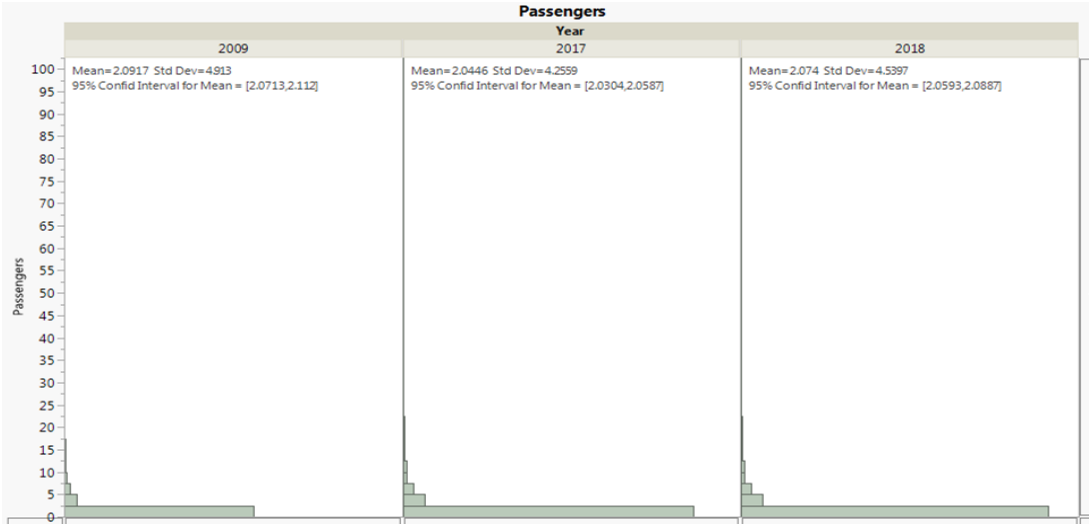

**Quantiles**

| | | |
|---|---|---|
| 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 |

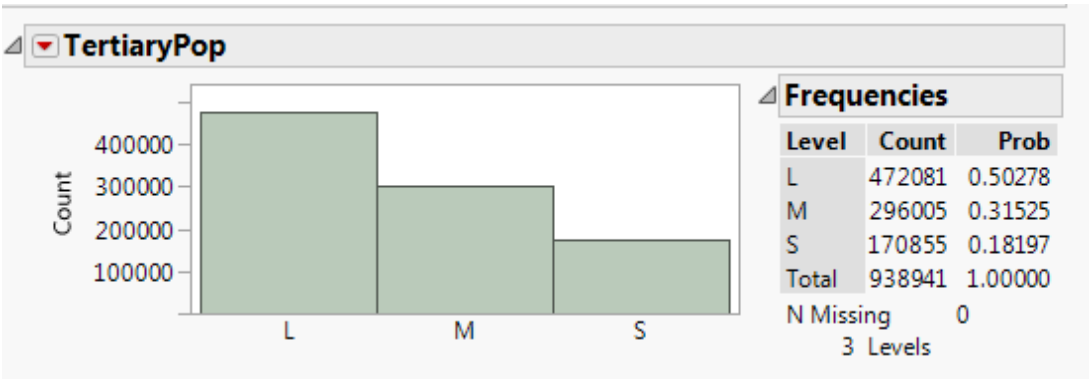
**Summary Statistics**

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

The average number of passengers for all three years combined was 2 per itinerary. 90% of the times the number of passengers was at most 3. The average number has been consistent at 2 for the years separately too.



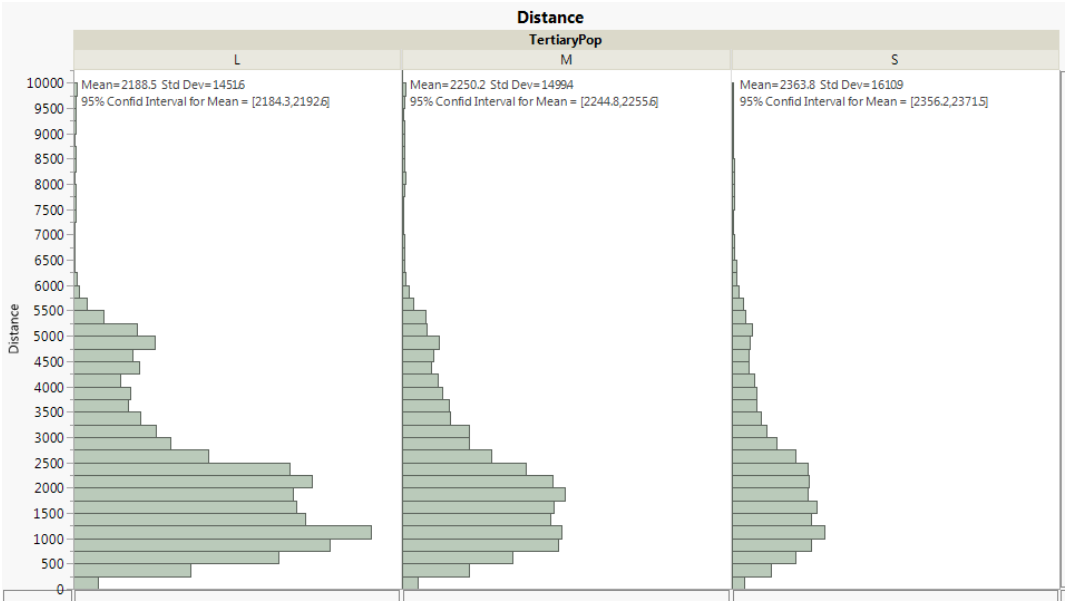

2. Create a new tertiary (three categories) flag variable using population "Pop" variable for states that have populations more than 10 million, 5-10 million and those less than that. Call this variable TertiaryPop (stands for Tertiary population).



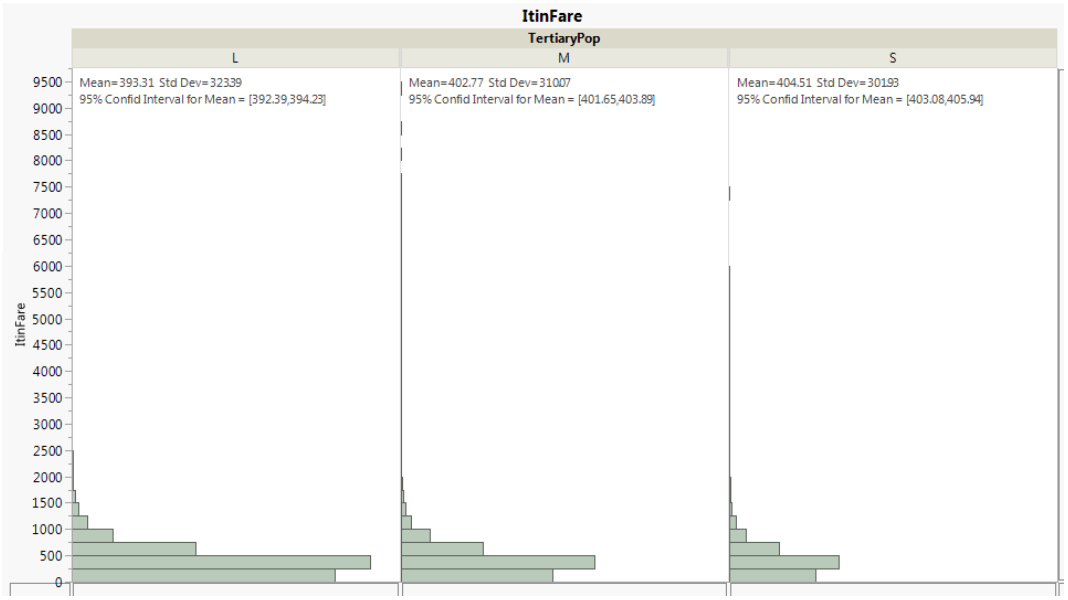

Pop > 10,000,000 –> L
5,000,000 < Pop ≤ 10,000,000 –> M
Pop ≤ 5,000,000 –> S

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-
The average distance travelled for large states is 2189 miles (95% CI at [2184, 2192]), medium states is 2250 miles (CI- [2245, 2256]) and small states is 2364 miles (CI- [2356, 2372]). Thus with increase in the size of the state, the average distance travelled has decreased.




ItinFare-




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]. Although the difference is small, the average ticket fare is inversely proportional to the size of the state- fare rises with decrease in state size.

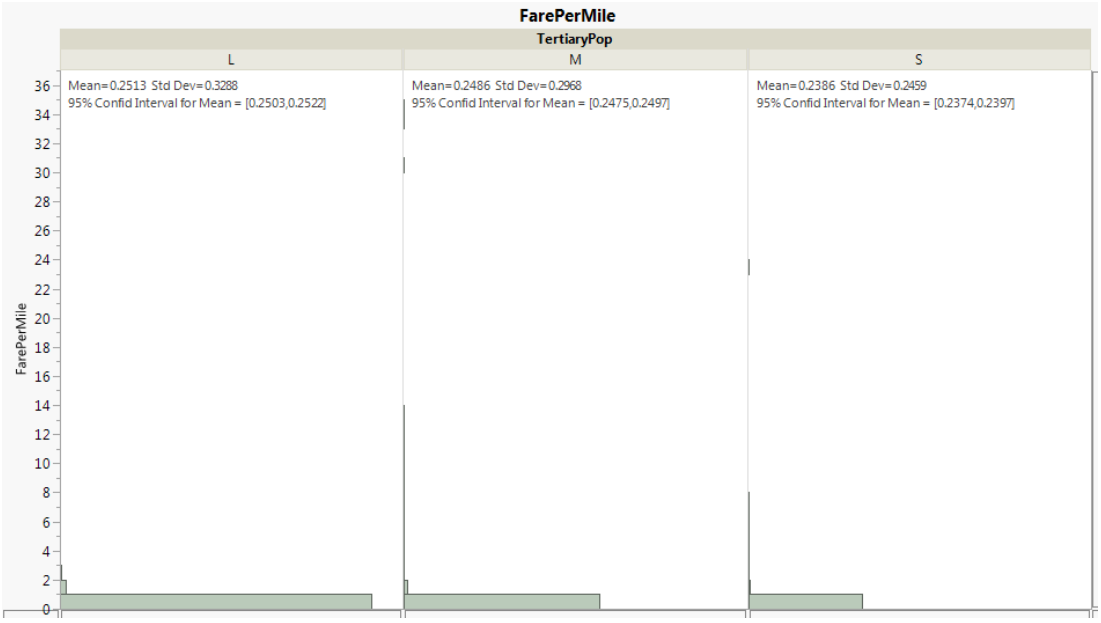
FarePerMile-

The difference in the fare per mile is very small in the three categories. Large, medium and small states had values at $0.2513, $0.2486, $0.2386. Although the value has decreased with size of state, it is very minuscule to be relevant.

CI for large states- [0.2503, 0.2522]

CI for medium states- [0.2475, 0.2497]

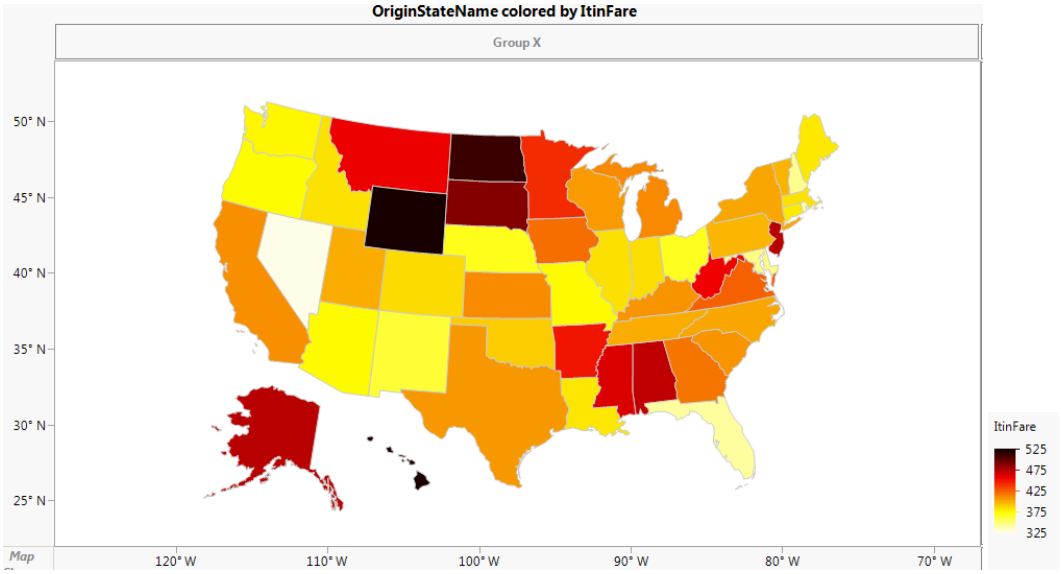
CI for small states- [0.2374, 0.2397]



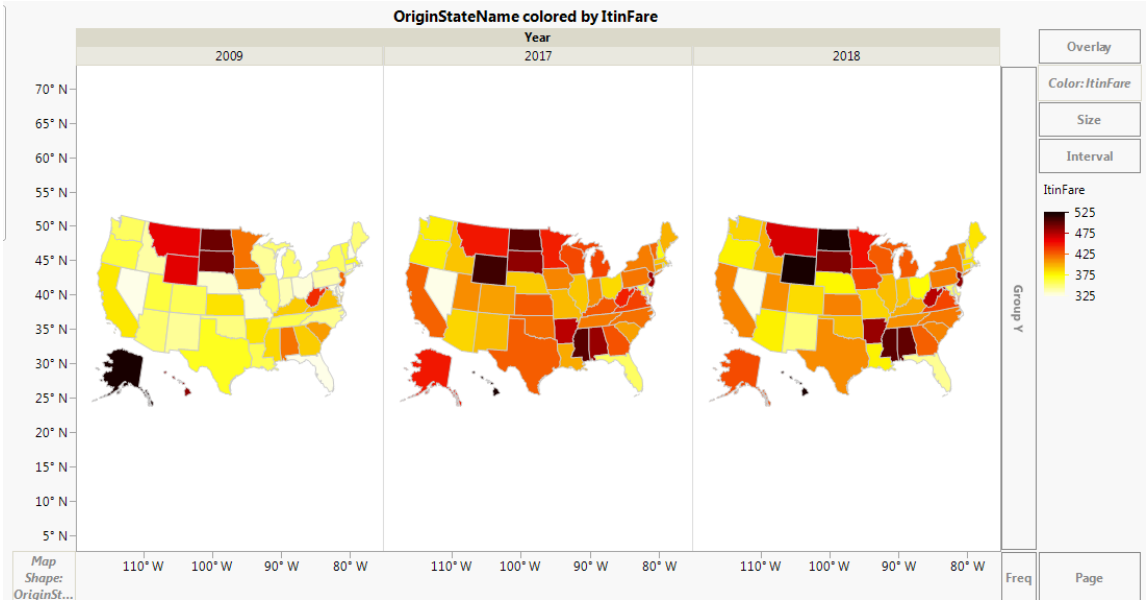

4. It is of interest for consumers to know if there are differences in airline ticket fares by the origin of the flight. In other words is there a difference in ticket fares to fly from different states? Run a comprehensive analysis of comparing average ticket price and average ticket price / mile by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

   ItinFare-

i. Wyoming is shown to have the highest ticket fare with $528 while Nevada is the cheapest with $321.

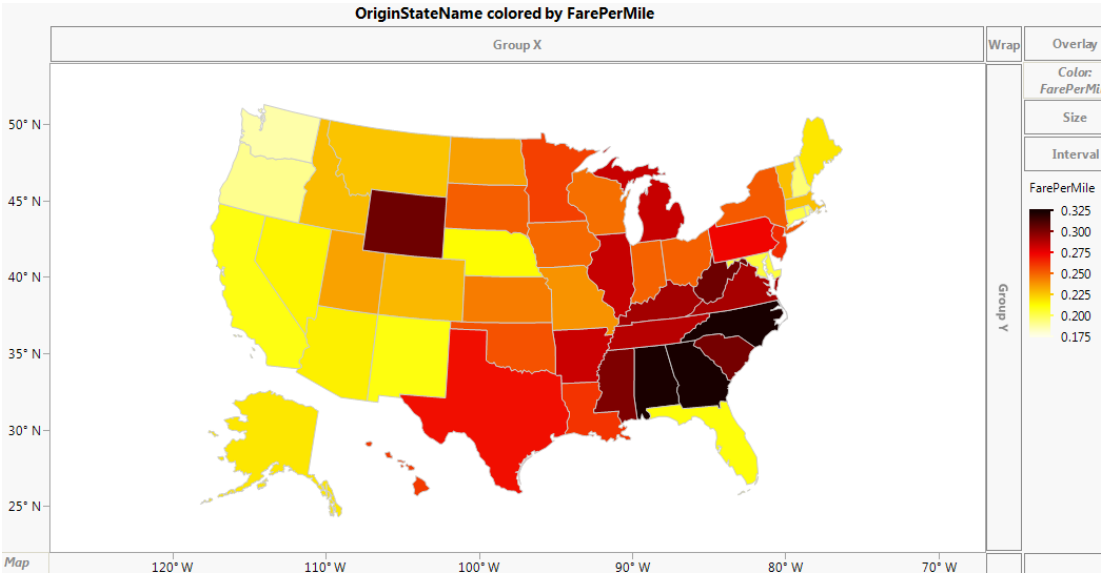


ii. In 2009, Alaska was the most expensive with $573 average ticket fare while Florida was the cheapest at $319. In 2017, Hawaii had the highest ticket rates at $535 while Nevada became the cheapest at $324. In 2018, Wyoming crossed Hawaii as the most expensive source of flight with average prices at $587 as Nevada remained as the cheapest at $318.



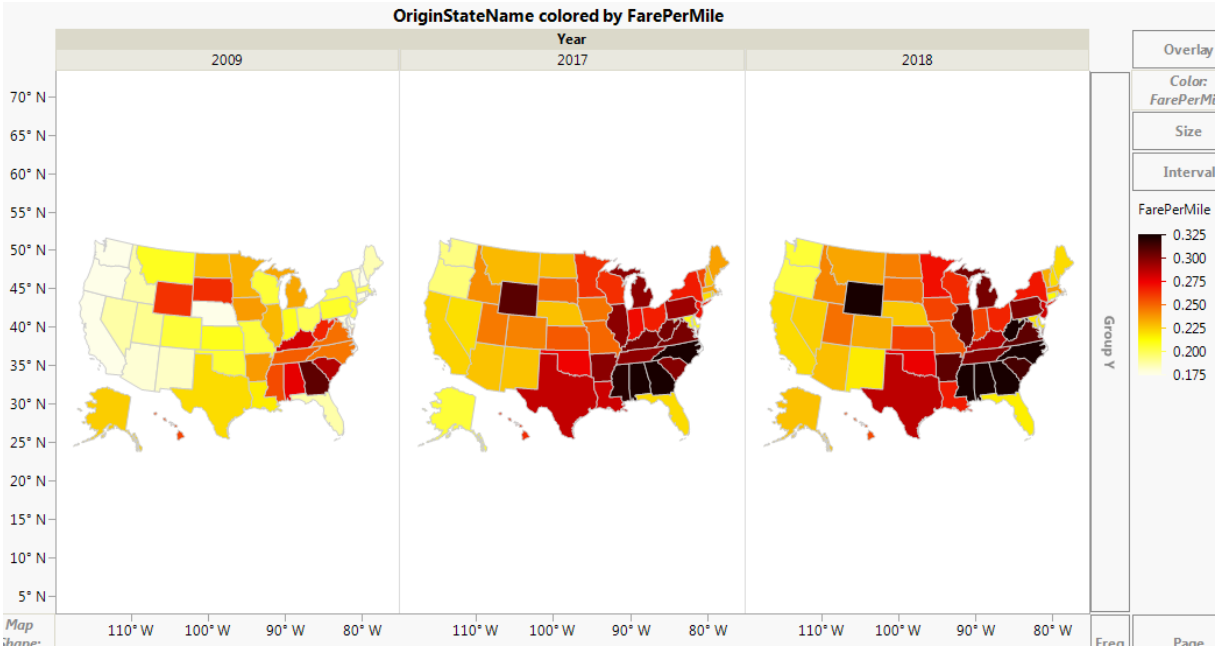

FarePerMile-

i.




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

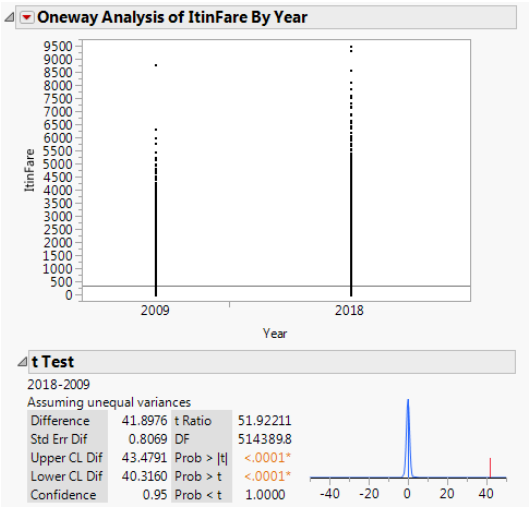
ii. In 2009, New Hampshire had the lowest fare per mile at $0.1448 as Georgia had the highest at $0.3090. In 2017, Washington State was the cheapest at $0.1932 while North Carolina topped with $0.3463. In 2018, California was the lowest $0.2024 and North Carolina remained at the top at $0.3583.




Bivariate Analysis

Run the following tests of hypothesis and provide relevant output and analysis in your own words.

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




Conditions-

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. $H_0: \mu_{2018} \leq \mu_{2009}$
   $H_1: \mu_{2018} > \mu_{2009}$

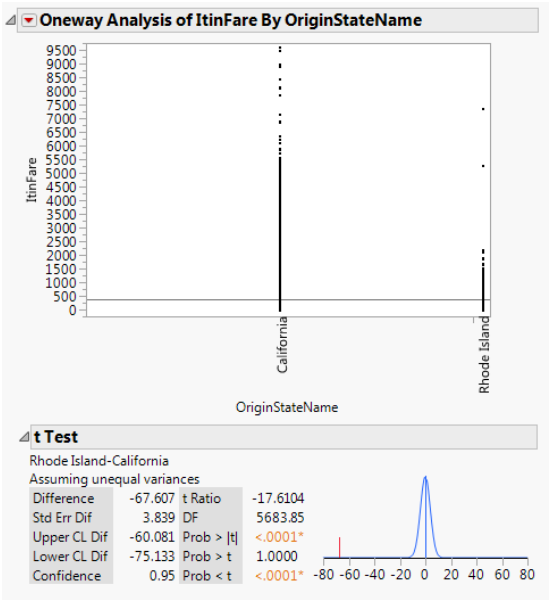
For Prob > |t|, the p-value is very less, so in (i), the null hypothesis that both the averages are equal can be rejected. Similarly, the null hypothesis that the average of 2018 is greater than

2009 can also be rejected. However, the p-value is very high for Prob < t, which means the null hypothesis that average of 2018 is lesser than that of 2009, cannot be rejected.

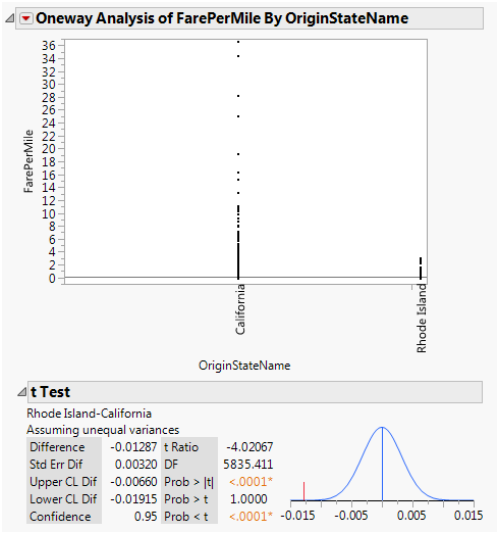
**This shows that the average itinerary fare in 2018 was higher than that of 2009.**

2. **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.

Average ticket prices-



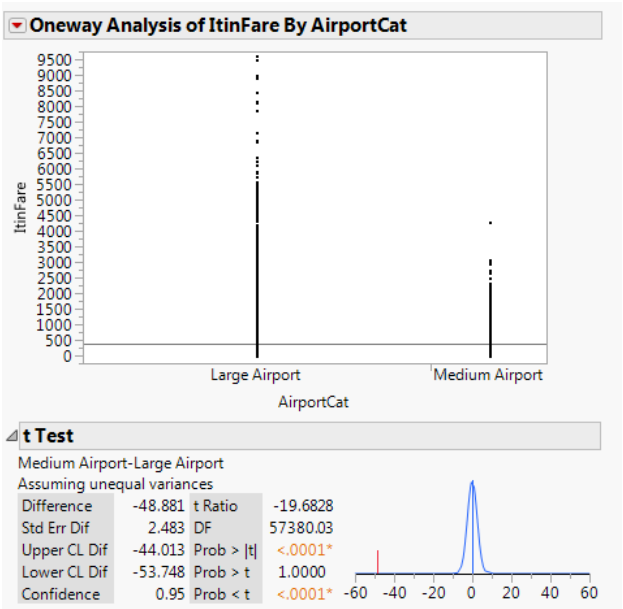

Here, the p-value for the hypothesis that the airfare from Rhode Island is less than or equal to that of California has a very high value. Hence the null hypothesis has failed to be rejected. However, the null hypotheses that the values are equal and that airfare from Rhode Island is greater than or equal to that of California have both p-values very less. This means that the null hypotheses can be rejected in favor of the alternate hypotheses that they are unequal and that airfare from Rhode Island is lesser than airfare from California. **Hence, it cannot be proven that it is more expensive to fly out from smaller states from the given values.**




The null hypothesis that the fare per mile to fly from Rhode Island is less than or equal to that of California cannot be rejected because of the high p-value. However, the p-value for the hypothesis that fare per mile from Rhode Island is greater than or equal to California is very

low along with the null hypothesis that they are both equal, thus proving that the value for Rhode Island is less than that of California. **Therefore, it cannot be proven that the fare per mile from smaller states is higher than that of larger states.**

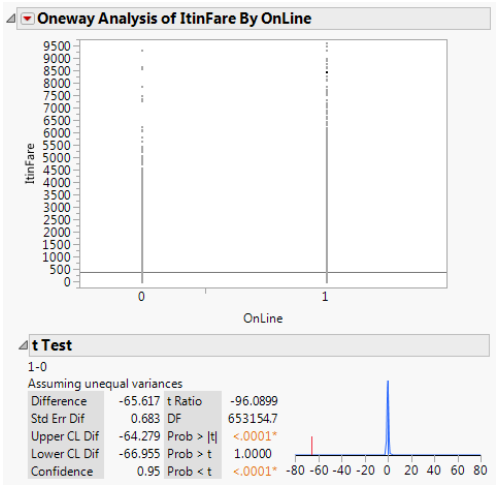
3. **Some airports within the same state are more expensive to fly out of-** State of California has ~ 30 airports with multiple airports present in larger cities. LAX, SAN, and SFO are three of the largest airports and SNA, SMF and OAK are three medium size airports. Combine the largest and medium size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



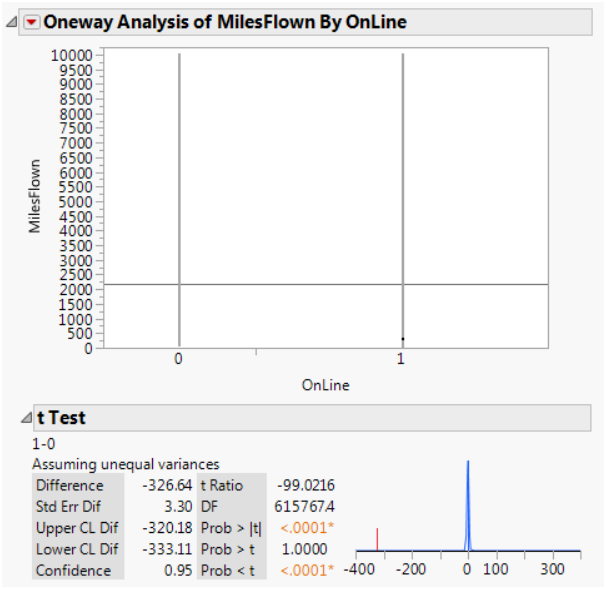

The null hypothesis that airfares from medium airports are less than or equal to airfares from large airports cannot be rejected due to the high p-value. Since the p-values are very low, the null hypotheses that the values are equal and that the airfares from medium airports are greater than or equal to airfares from larger airports can be rejected in favour of the hypotheses that they are unequal and that airfares from medium airports are cheaper than that of large airports, **thus proving that large airports from California are expensive to fly out from compared to medium airports.**

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-
   Since the p-value is very high, the null hypothesis that online tickets are cheaper than or equal to tickets purchased offline cannot be rejected. However, hypotheses that online tickets are more expensive or equal to offline tickets can be rejected due to very low p-value. This means that the alternate hypothesis that **online tickets are cheaper than offline tickets** is true.




Miles Flown-

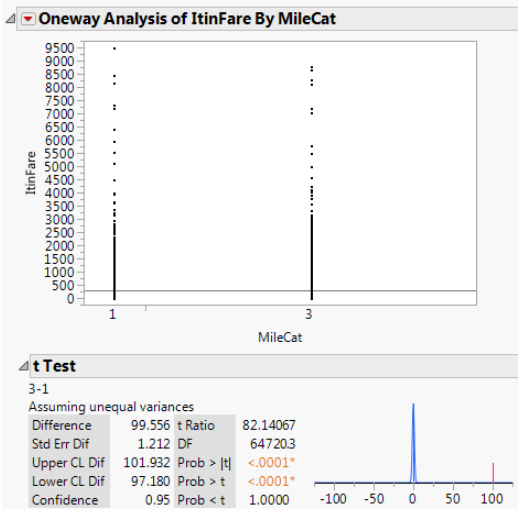



The p-value for the null hypothesis that online tickets have higher or equal miles as that of tickets purchased offline is very low. So the null hypothesis can be rejected in favor of the hypothesis that **online tickets have lower miles flown compared to offline tickets**.

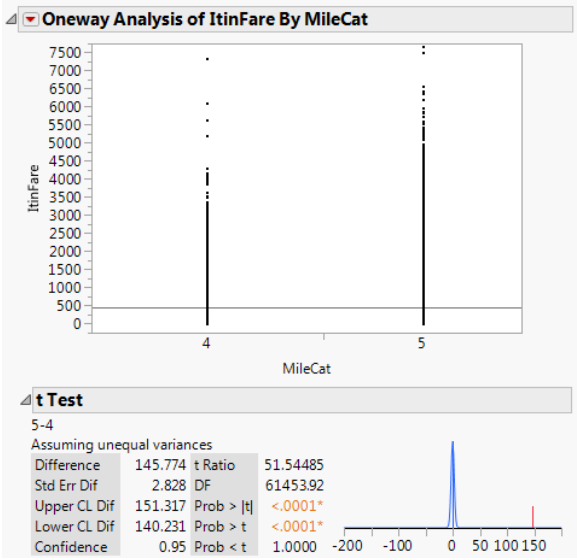
5. **The longer the flight, the pricier it is-** Test this theory by creating a four category variable from MilesFlown variable: "< 500 miles", "500 – 1200 miles", "1200 – 2000 miles", "2000 - 3000 miles", "3000+ miles". Run the following tests of hypothesis:
    a. Is there a difference in average fare cost between flights that flew "<500 miles" and those that flew "1200-2000 miles"?
    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.
    c. 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

I have assigned the following values for the categories-
<500 –> 1; 500-1200 –> 2; 1200-2000 –> 3; 2000-3000 –> 4; >3000 ->5

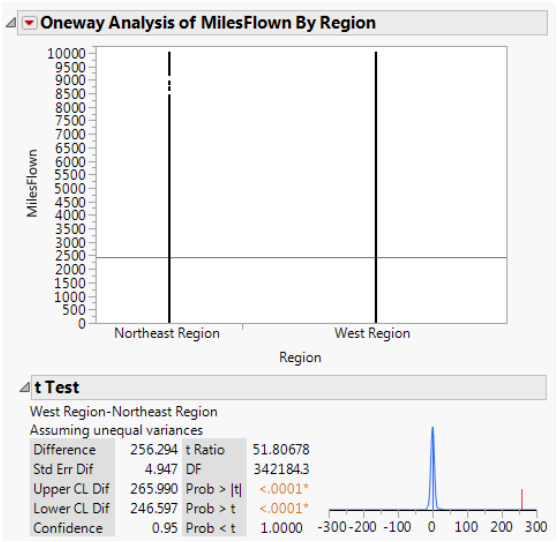
a. From the t-test analysis of ItinFare based on the two categories, the p-value for the null hypothesis that the average airfare for flights that flew less than 500 miles is greater than or equal to those that flew between 1200 and 2000 miles is very low. Hence, this hypothesis can be rejected in favor of the hypothesis **that the average airfare for flights that flew between less than 500 miles is less than those that flew between 1200 and 2000 miles.**



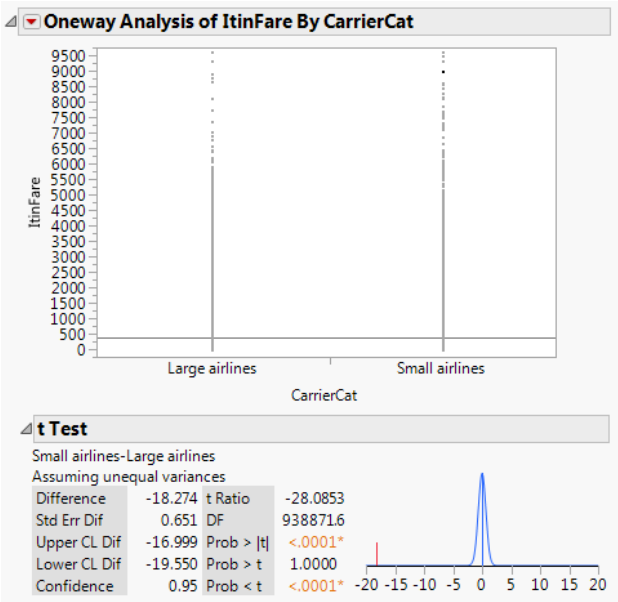

b. The p-value for the null hypothesis that average fare cost between flights that flew between 2000 and 3000 miles is greater than or equal to that of 3000+ miles in the Northeast region is very low. Hence, this hypothesis can be rejected in favour of the hypothesis **that average fare cost between flights that flew "2000 – 3000 miles" is less than "3000+ miles" in the Northeast region.**




c. 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 very less. Hence, it proves that **miles flown from airports in the Northeast region is lesser compared to the Western region.**




6. **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.




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