<https://dasl.datadescription.com/datafile/flights-on-time-2016/>

<https://drive.google.com/open?id=1KrOU3T9T4AbOxDJzEqmfI4ZcarKZLI7i58sKzby5A5E>

<https://www.transtats.bts.gov/HomeDrillChart.asp?URL_SelectYear=2016&URL_SelectMonth=6&URL_Time=1&URL_Selection=1>

Let's brainstorm with the new dataset:

proc means data=webwork.flights mean skew std cv skew stderr;

var cancelled\_pct;

class month;

Run;

Look at the CV of the last month, september. By confirmation with the calendar there are several celebrations in the US which

I am not understanding what the columns year means.

What to include in Handout for Presentation

* Summarize purpose
* Variables used
* Statistical analysis
* Conclusion summary

Outline for Paper - APA format

* Abstract (no more than 200, often less than 100 words)
  + General subject of research, scope of research, identification of type of methodology used, conclusions
  + Project’s objective
* Background and Purpose (at least 5 references)
  + Importance
    - “Flight delay is a serious and widespread problem in the United States. Increasing flight delays place a significant strain on the US air travel system and cost airlines, passengers, and society at many billions of dollars each year” (Zou et al., 2010, 1)
    - Build into Marketing
      * “When considering an airline, travelers also look for timelines and assurance. The 2016 Travel Report from American Customer Satisfaction Index listed on-time arrival as one of the major elements of the customer experience that determine satisfaction” (Dyer, 2016)
    - Offer on-time performance commitments
      * “Airlines often use on-time performance ranking to offer service level agreements to corporations that greatly influence purchasing decisions” (Dyer, 2016)
    - Analyzing on-time performance from a third party is important
      * Removes skepticism from when an airline releases analyses, data about own operations
      * Provides credibility
  + Literature review- we each should find at least 1 other study
    - Zou et al., 2010
      * Other studies have attempted to estimate total economic impact of delays
        + This report analyzes cost components

Cost to airlines, cost to passengers, cost of lost demand, indirect impact of delay on US economy

* + - * + Use delay based Tanslog cost functions models
        + “Of particular interest to this study are the delay variables, the estimates of which support our hypothesis that excessive flight time affects airline costs” (Zou et al., 2010)
        + Capacity Induced Schedule Delay: Airlines tend to generate “peaked schedules”, capacity constraints may force flights to move to a less congested time of day. Generalized linear regression used to “handle auto-correlation in the residuals with different moving average lag periods” (Zou et al., 2010,45)
        + Value of Demand Lost Due to Delays: Found that “there could be differences between the average flight delay and average delay experienced by each passenger on the route. Nonetheless high correlations exist between the two variable” (Zou et al., 49)
* Sample Collection Techniques, methodology, design
  + Data collection process
  + How and when you acquired data
  + Size of sample and information about variables of interest
  + Information about margin of error for percentages quoted in report based on sample sizes used
  + Describe design (if applicable, power of the test)
  + Statement of research
    - Set of research hypotheses
      * This will determine what tests and methodology that will be needed
    - Statement of research hypotheses can be included as part of purpose (or after descriptive statistics as part of body of your report)
* List of variables, responses, responses and codebook of variables in analysis
* Routines in SAS
* Body
  + Tabular and numeric descriptive statistical techniques
    - Supported by tables or graphs
    - Summary of descriptive findings
    - Include margin of error for all percentages quoted
  + Summarizes inferential procedures (regression/correlation/t test/analysis of variance)
    - Correspond to research hypotheses previously stated
    - Rationale for using certain procedures
    - Include hypotheses, assumptions, test statistic, p value and conclusion
    - Supported by tables/ graphs
* Summary/Conclusion and Recommendation
  + Summary of findings
  + Conclusion then outcomes and decisions based on research
  + Recommendations are suggestions for how to proceed based on conclusion
  + List limitations
    - How did they compare with research hypotheses
    - What other variables would you have tried to include in study
    - Were there variables you would not use in further studies
  + Suggestions for future research
* References
  + Zou, B., & Hansen, M. (2010). Flight delay impact on airfare and flight frequency: A comprehensive assessment. *Transportation Research Part E: Logistics and Transportation Review,* *69*, 1-94. doi:10.1016/j.tre.2014.05.016
  + Dyer, K. (2016, December 15). Why is on time performance important to airlines. Retrieved March 5, 2019, from https://www.flightstats.com/company/why-is-on-time-performance-important-to-airlines/
  + U. S. Department of Transportation (2016). Flights on time 2016. Retrieved March 1, 2019, from <https://dasl.datadescription.com/datafile/flights-on-time-2016/>
* Appendix

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Reference Source and some research

<https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20070038363.pdf>

Feature extraction techniques attempt to aggregate the predictors in some way in order to extract the common information contained in them that is most useful for model building. Typical methods include Factor Analysis and Principal Components Analysis, Multidimensional Scaling, Partial Least Squares methods, and others.

<https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_factor_sect028.htm>

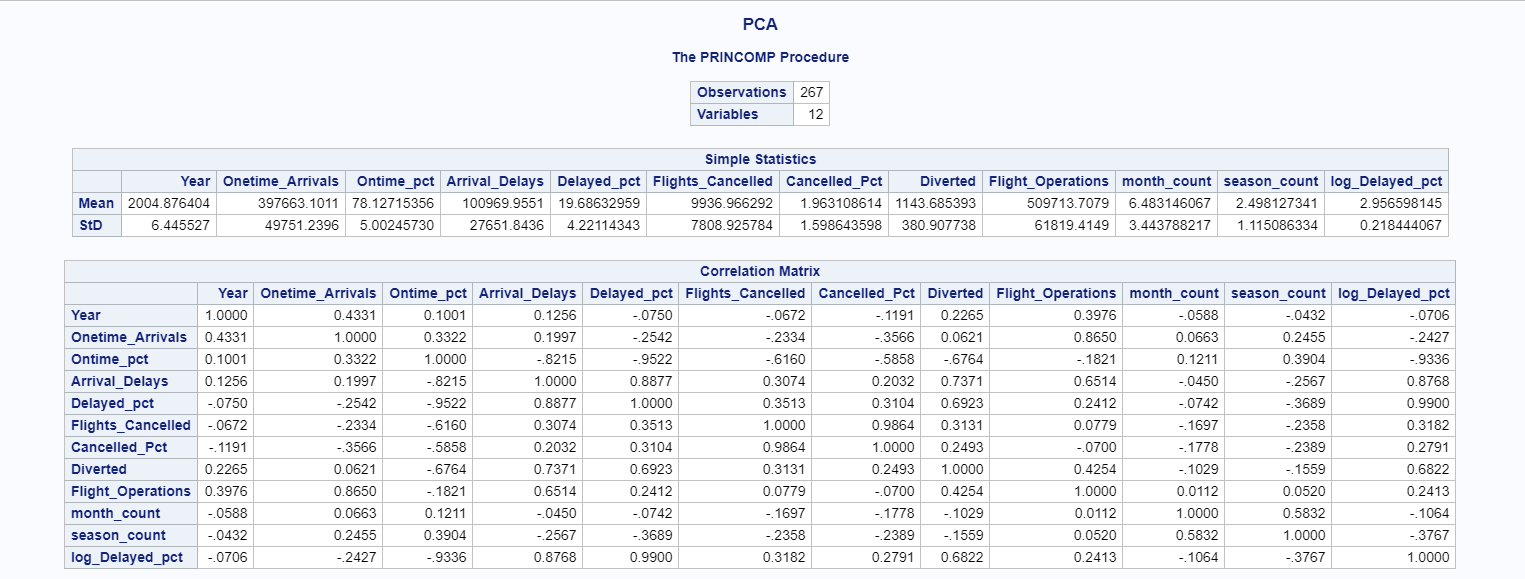
Advanced techniques used to find relationships between multiple variables to discover the variance.

Example:

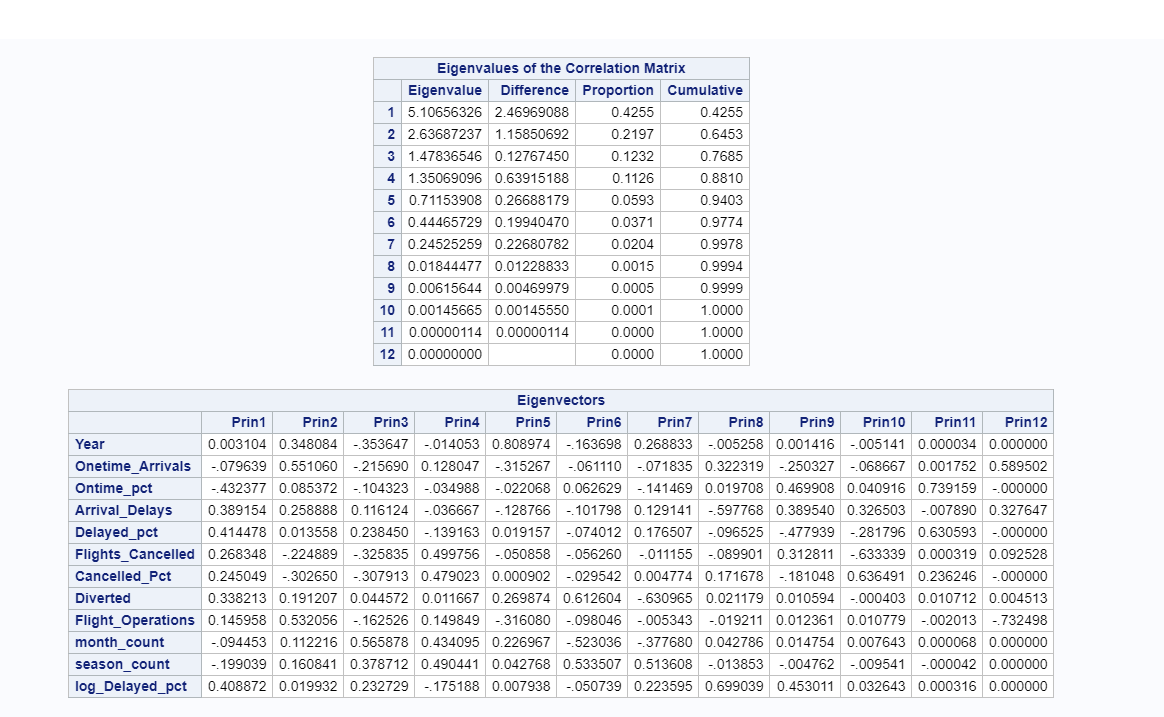
proc princomp data=new plots=(Matrix PatternProfile);

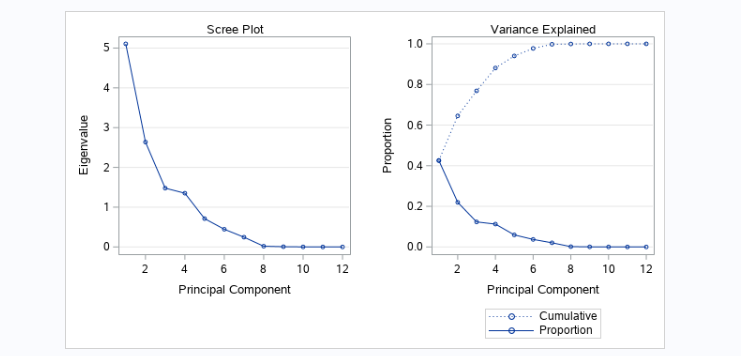
title 'PCA';

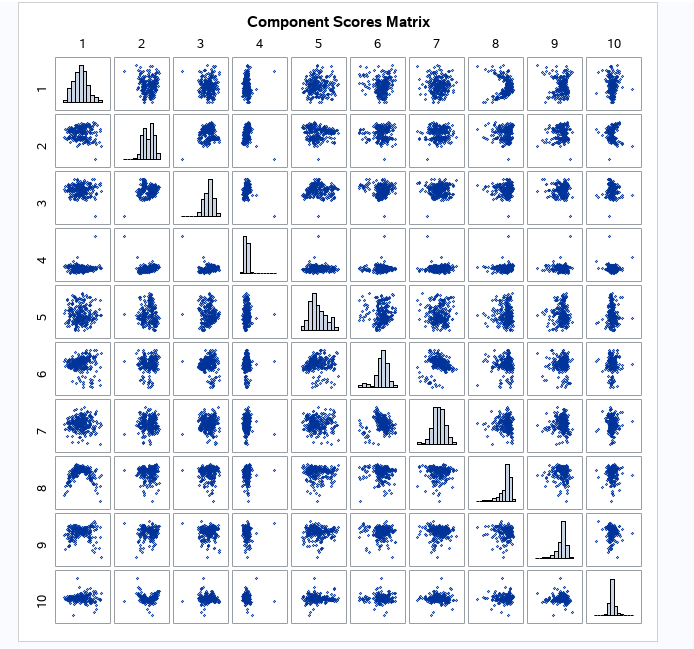
run;

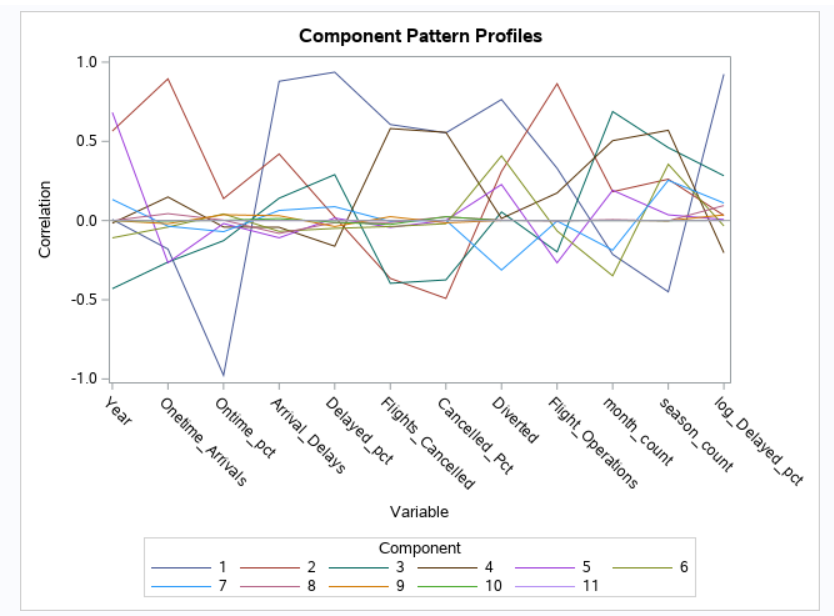


PROC PRINCOMP displays the scoring coefficients as eigenvectors, which are shown in. PROC PRINCOMP creates components that have variances equal to the eigenvalues.









####################################################################################

Analysis of Variance on Arrival delays using GLM

Find what is the difference between Arrival delays each in older years before 2005 compare to later years After 2005

First we will run the glm on variable Arrival Delays , then if we find the difference in distributions we

Will run the glm on Flight Operations to find out if Difference in distributions of flight operations in older

Years was significantly different compared to new years.Than if they are significantly different we will

Run the GLM on Delay pct to make the final decision.

proc glm data=new ORDER = DATA ;

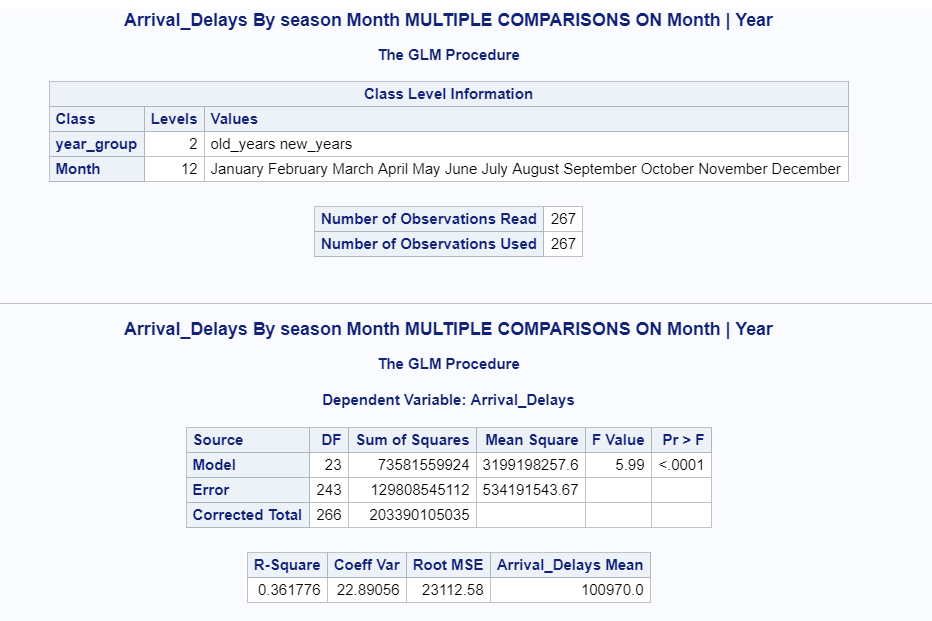
title 'Arrival\_Delays By season Month MULTIPLE COMPARISONS ON Month | Year';

class year\_group Month;

model Arrival\_Delays=year\_group | Month;

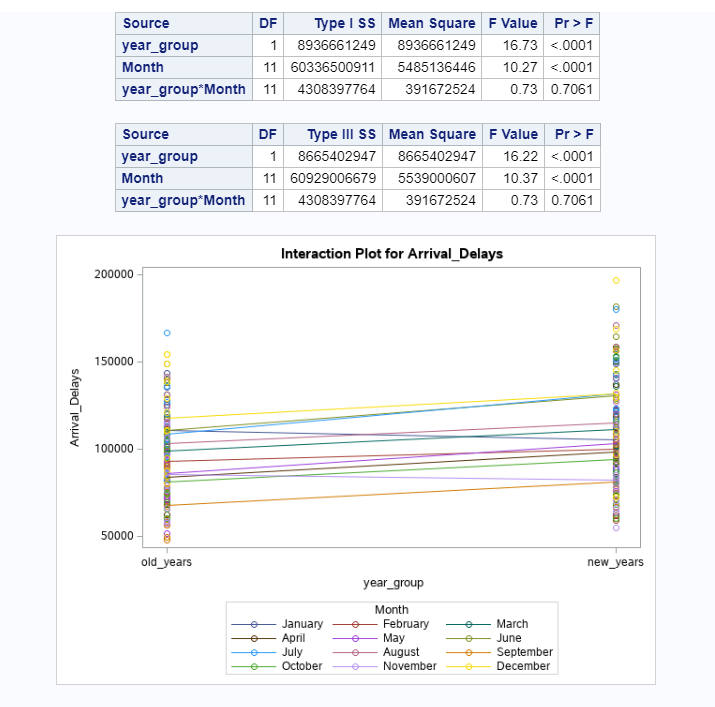
means year\_group | Month / snk;

run;



F-Value =5.99 p-value < 0.001

This p value shows how well the model displays the variation about the grand mean.



Ho -> The difference of Arrival delays distributions in older years compare to later years are not significant

Ha -> The difference of Arrival delays distributions in older years compare to later years are significant

Ho -> The difference of Arrival delays distributions for each month are not significantly different.

Ha -> The difference of Arrival delays distributions for each month are significantly different.

Ho -> The difference of Arrival delays distributions for month and year group interactions are not significantly different.

Ha -> The difference of Arrival delays distributions for month and year group interactions different.

For year group: F value = 16.73 p-value=0.0001 < alpha=0.05

For month group F value=10.27 p-value=0.0001 < alpha=0.05

For year\_group\*Month interaction F value=0.73 p-value=0.7061.alpha=0.05

Reject Ho for both years and month

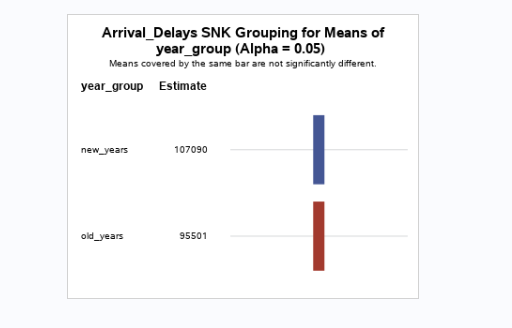
Fail to reject for year group and month interactions

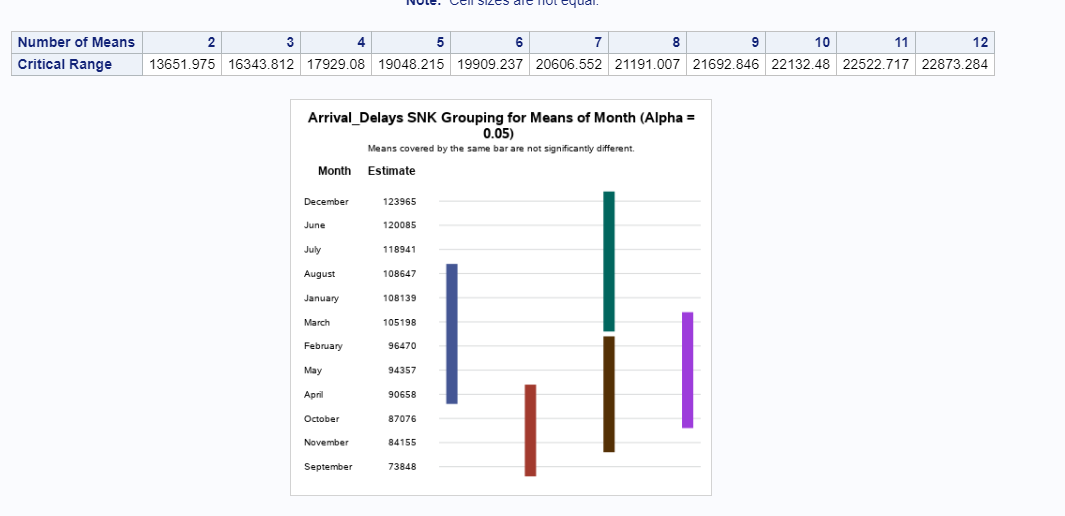
Conclusion :

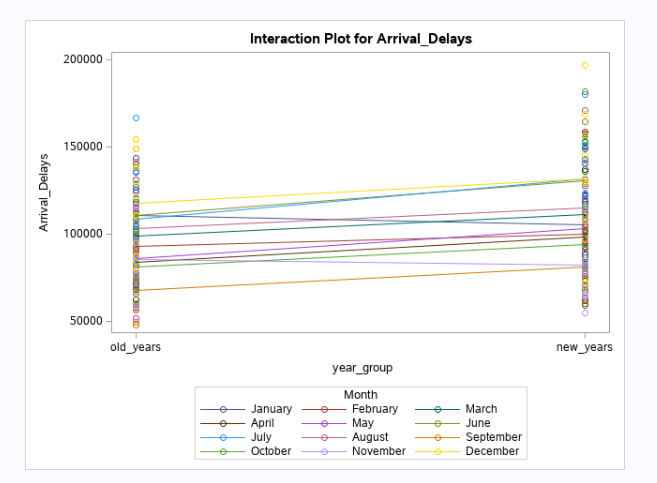
At significance level alpha=0.05 the distributions of Arrival delays for each month are significantly different.

At significance level alpha=0.05 the distributions of Arrival delays for older years compare to later years are significantly different.

At significance level alpha=0.05 the distributions of Arrival delays for year group and month interactions are significantly different.







But now we have to compare the flight operations. If the flight operations distributions are the same

in older years compare to later years then there is a significant difference between Arrival Delays.

I other case we have compare Delay pct.

**Flight Operations**

**proc glm data=new ORDER = DATA ;**

**title 'Flight\_Operations By season Month MULTIPLE COMPARISONS ON Month | Year';**

**class year\_group Month;**

**model Flight\_Operations=year\_group | Month;**

**means year\_group | Month / snk;**

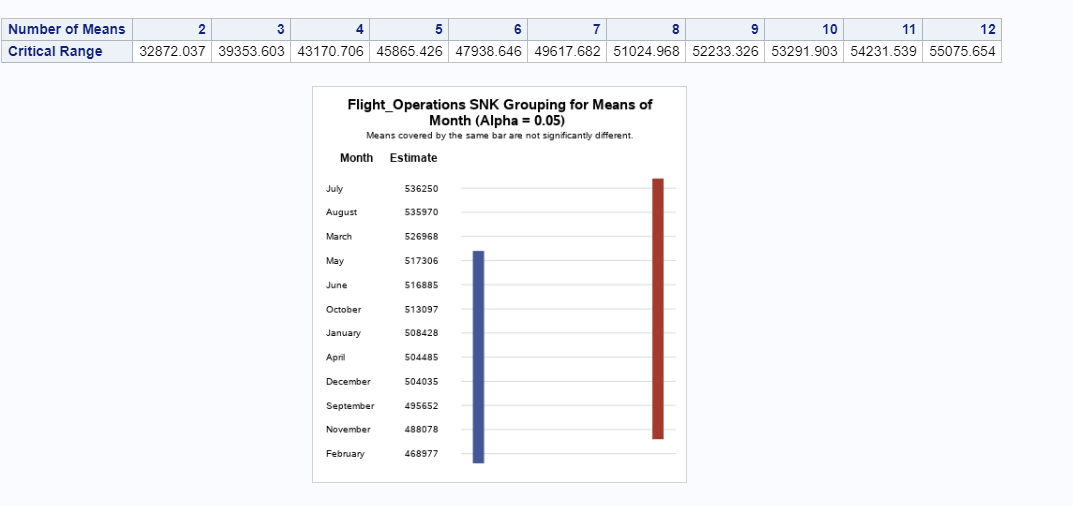
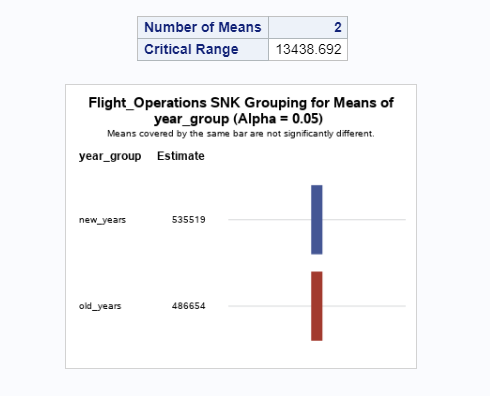
**run;**

****

**We can see that for flight operations distributions are significantly different for both**

**Groups month and the years but not significant for their interactions.**

**The plot shows the in laters years the flight operations are higher that in older years.**

****

Delay pct

proc glm data=new ORDER = DATA ;

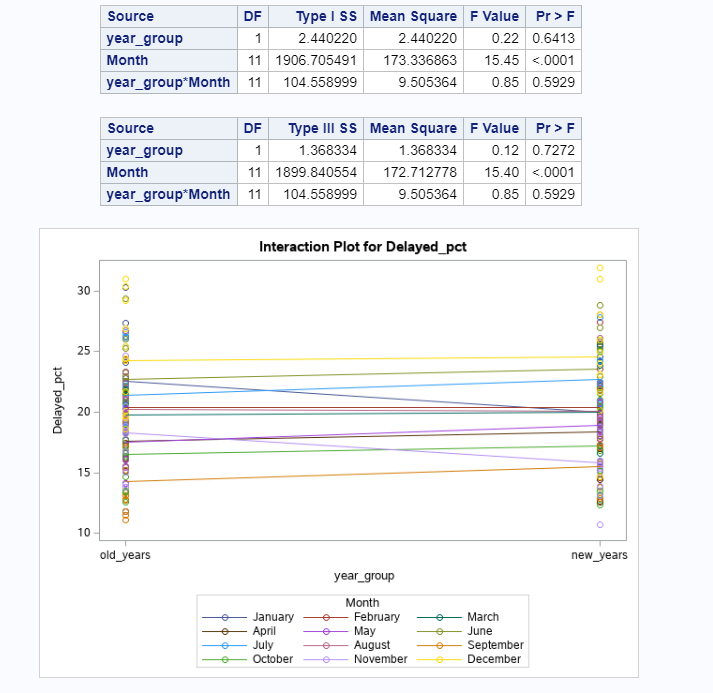
title 'Delay pct By season Month MULTIPLE COMPARISONS ON Month | Year';

class year\_group Month;

model Delayed\_pct=year\_group | Month;

means year\_group | Month / snk;

Run;



Ho -> The difference in distributions for Delay pct in older years compare to later years are not significant

Ha -> The difference in distributions for Depay pct in older years compare to later years are significant

Ho -> The difference in distributions for Delay pct for each month are not significant

Ha -> The difference in distributions for Delay pct for each month are significant

Ho -> The difference in distributions for Delay pct for interactions of month and year group are not significant

Ha -> The difference in distributions for Delay pct for interactions of month and year group are not significant

For year group: F value = 0.22 p-value=0.6413 < alpha=0.05

For month group F value=15.45 p-value=0.0001 < alpha=0.05

For year\_group\*Month interaction F value=0.85 p-value=0.5929.alpha=0.05

Reject Ho for month group

Fail to reject for year and interactions

Fail to reject for year group and month interactions

Conclusion :

At significance level alpha=0.05 the difference in distributions of Delay pct at least for one month are different for one of other.

At significance level alpha=0.05 the difference in distributions of year group in older years compare to later years are not significant.

At significance level alpha=0.05 the difference in distributions of year group and month interactions are not significant.