FAIRLEIGH DICKINSON UNIVERSITY

DSCI 6780 Final Project

Running a Six Sigma Project in Amazon Transportation Services

Due: 6pm December 18 (Monday), 2023. Please submit on time on WebCampus in order to meet FDU Grade Report Deadline.

FAIRLEIGH DICKINSON UNIVERSITY

COVER PAGE

The data used in DSCI 6780 final project are all real data from Amazon.com. Therefore, students shall agree **NOT** to disclose the data to any third party, individuals, and organizations. Students shall **NOT** use the data for any other commercial activities other than the scope of DSCI6780 assignment.

Documents to Submit on WebCampus:

1. **Main Document (Only the main document will be graded)**
   This word document outlining your answers for each question with relevant graphs, tables, regression outputs, etc.
   - Excel and Python files are only supporting documents.
   - Readers should NOT open the supporting documents to find answers.
   - Your word document should include all details of your work from Excel and Python.
2. **Supporting Documents (Supporting documents will not be graded):**
   1) One single **Excel** file for the Define stage.
   2) One single **Excel** file for Simulation in Improve stage.
   3) One single **Python** file in *pdf format* for all the other questions. Include all Python outputs. Use headings to separate different sections of the questions.

Amazon transportation services is the transportation arm of Amazon that handles U.S. domestic deliveries using Amazon-owned trailers and 3rd party trailers. The ideal state is that when the planning is done for each truckload, the delivery will happen based on what has been planned. However, due to uncertainty in customer demand, demand fluctuations are challenging to predict. Therefore, Amazon has to frequently reposition its available trailers on daily basis based on demand fluctuations, this is referred to as Trailer Pool Adjustment (TPA) in Amazon. TPA is considered a waste in the operations process as it creates extra costs by running empty trailers around. Our major task in this project is to use Six Sigma knowledge to help improve the trailer pool adjustment activities as well as to improve the planning accuracy.

Define in Excel (5%):

A clean file named '*TPA.xlsx'* is available. This file consists of information of TPA movement in 62 Tier 1 Amazon warehouses from Jan 1, 2020 to Sep 30, 2020 on daily basis.

Definition of variables is given below.

- Dom: domicile. This refers to the 62 Tier 1 Amazon warehouses.
- loadout: how many truckloads coming out of a warehouse each day.
- loadin: how many truckloads going into a warehouse each day.
- loadTotal: inbound + outbound truckloads each day
- loadingtime: time spent loading shipments onto trailers.
- unloadingtime: time spent unloading shipments from trailers.
- ntrailers: number of available empty trailers to be used for loading at each warehouse.
- tt_time: average transit time for each warehouse.

- TPAIn: inbound TPAs (how many additional empty trailers a warehouse has requested to be sent *in*).
- TPAOut: outbound TPAs (how many empty trailers a warehouse has sent *out* to support other warehouses)
- TPATotal: inbound TPAs + outbound TPAs

Task:

Do a Pareto analysis for the column of *TPATotal* and answer the questions:

1. How many warehouses represent 50% of cumulative TPAs?
2. How many warehouses represent 80% of cumulative TPAs?
3. Is this a typical Pareto analysis?
4. What are the top 5 warehouses you can focus on to start improving the process?

Measure in Python (15%):

You want to do further analysis to find out what factors contribute to TPA movement. You also want to highlight the cost of TPA in the top ranked warehouses to the management team.

To continue your analysis, you want to make sure your data is clean.

Task 1:

1. Use Python to detect if there are any outliers in the three columns of '*loadingtime*', '*unloadingtime*', and '*tt_time*'. Please report your results and show the outlier graphs generated by Python.
2. If you found any outliers in the three variables of '*loadingtime*', '*unloadingtime*', and '*tt_time*', remove those outliers from your data (meaning *the entire row will be removed if there is an outlier in it*). Then, use Python to conduct a Normality test. Do these three variables look normally distributed after you remove the outliers? Report your result along with the graph generated by Python.

Task 2:

Two files are available: '*ONT.xlsx'* and '*DFW.xlsx*'. You want to highlight the TPA cost to the management team in these two warehouses. The column of '*total_est_cost'* is the estimated TPA cost for the two warehouses of ONT and DFW.

1. Use Python to detect if there are any outliers in the column of '*total_est_cost*'.
2. If you found any outliers, remove those outliers from your data (meaning *the entire row will be removed if there is an outlier in it*). Then, use Python to plot the Histograms of TPA cost for ONT and DFW. Place the two Histograms side by side in Python and report your result. Did you observe anything abnormal from the histogram? Does the cost make sense? Any suggestions to re-plot the histogram to make more business sense?
   Hint: Remember to choose your bin range properly for the two warehouses.
3. Based on your recommendations and thoughts from Step 2, plot two new Histograms of TPA cost for ONT and DFW. Place the two Histograms side by side in Python and report your result.

Analyze using Python (25%):

You want to establish what factors contribute to TPA movement. You continue using the file of '*TPA.xlsx*' to do a regression analysis (**after** removing potential outliers from the columns of loadingtime, unloadingtime, and tt_time).

Task 1:

1. Use '*TPA.xlsx*' to run a regression analysis in Python. Your y variable is TPATotal. Your x variables are: loadout, loadin, loadTotal, loadingtime, unloadingtime, ntrailers, tt_time.
2. What are the factors that impact TPA movement? (Hint: check the *p* value for each coefficient. If a coefficient has a *p* value ≤ 0.05, it is considered statistically significant, meaning the corresponding variable contributes to TPA movement). Briefly discuss your model fit and your results.

Task 2:

You suspect that time spent in loading and unloading shipments at each warehouse may play an important role in TPA movement. A new set of data is available '*train.csv*' (Assume all data in '*train.csv*' are correct and no outliers. No need to detect and remove outliers). Definition of variables is given below:

- sin_departure_time_hour_of_day: departure hour of day converted to sine.
- cos_departure_time_hour_of_day: departure hour of day converted to cosine.
- departure_hour_of_day: trailer departure hour in a day
- departure_day_of_week: trailer departure day in a week
- departure_week_of_year: trailer departure week in a year
- hook_trailer_min: how much time spent in loading
- drop_trailer_min; how much time spent in unloading
- average_transit_hour: average transit time between two warehouses
- miles: distance between two warehouses
- checkin_time_windows_at_origin: time window of driver check-in at origin
- total_block_minutes: system available time in mins to plan trailer activities
- num_feasible_blocks: system available time blocks to plan trailer activities
- origin_zip3: 3 digit of origin zip
- dest_zip3: 3 digit of destination zip
- lead_time_to_departure: time given to drivers to finish paperwork at origin
- is_eventually_unplanned: was the truckload being *unplanned into* Amazon planning system? 1 is Yes (meaning unplanned). 0 is No (meaning planned).

Your task here is to test two simple hypotheses at 0.05 level. For this task, use the easy approach in Python to conduct hypothesis testing.

1. Mean time to hook trailers is 25 min.
2. Mean time to drop trailers is 5 min.

Improve using Python and Excel (35%):

Here we have two tasks.

Task 1:

You want to help improve the planning accuracy to avoid unnecessary trailer movement. Two files are to be used: '*train.csv*' and '*test.csv*'. In '*train.csv*' and '*test.csv*', each row represents one truckload and its associated features. Assume all data in '*train.csv*' and '*test.csv*' are correct and no outliers. No need to detect and remove outliers in these two files.

1. Build a logistic regression model using '*train.csv*'.
   Hint:
   Column 'is_eventually_unplanned' is y and all the other columns are X.
   If Python gives an error message of "TOTAL NO. of ITERATIONS REACHED LIMIT", make sure you set max_iter to a larger number in your model (refer to class Python script for this).
2. Report confusion matrix of your model. Briefly discuss the meaning of your confusion matrix.
3. Report precision, recall, and accuracy measures. Briefly discuss the meaning of each measure.
4. Plot ROC Curve for your model. Briefly discuss the ROC Curve.
5. Using the model built from step 1 to predict the probability of being *unplanned* for the truckloads in '*test.csv*'. How many truckloads are predicted to have a more than 50% probability of being ***unplanned***?
   Hint: All columns in test.csv are your new X.

Task 2:

The operations manager at ONT wants you to do a simulation to tell him on average, how many trailers should ONT keep onsite on daily basis. Use the file '*TPA.xlsx*' (after removing potential outliers from the columns of loadingtime, unloadingtime, and tt_time). Relevant columns for this task are dom and ntrailers.

1. Create a frequency table to summarize the daily frequency of trailers for ONT. Use a bin range at a 50 interval from 50 to 450.
2. Once you have created the frequency table, you can start building a Monte Carlo simulation in Excel. Simulate a single run of 365 days with a total run of 100 times. Report the average outcome of the 100 runs.
3. What is your recommendation to the operations manager at ONT?

Control using Python (20%):

Use Python to answer the following questions:

Task 1:

Use '*train.csv*'. Assume all data in '*train.csv*' are correct and no outliers. No need to detect and remove outliers.

The required hook trailer time is 25±5 minutes, and the required drop trailer time is 5±1 minutes. Use Python to answer the following questions:

1. What is the Process capability Ratio (Cp) for hook trailer time and drop trailer time?
2. What is the Process capability Index (Cpk) for hook trailer time and drop trailer time?
3. Use the reference value of 1.33 to determine the capability of the process. What can we say about the hook trailer time and drop trailer time based on the data you have?

Task 2:

Use the original file '*TPA.xlsx'* again (**without** removing any outliers from the columns of loadingtime, unloadingtime, and tt_time). Calculate the average loading time (column '*loadingtime*') and average unloading time (column '*unloadingtime*') for the top 5 warehouses at monthly level (top 5 warehouses from your answer of Q4 in Define stage). You can use Excel Pivot to handle the data manipulation (a new column "month" needs to be created for pivoting purpose). You table will be looking like this:

Top Warehouses	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
Warehouse1									
Warehouse2									
Warehouse3									
Warehouse4									
Warehouse5									

Read in your table into Python and use Python to answer the following questions:

Loading Time:

1. What is the R bar ($\bar{R}$) value and X double bar ($\bar{\bar{X}}$) value for your sample?
2. What is the UCL and LCL of the R-Chart and X-bar Chart?
3. Plot R bar chart and X bar chart in Python.
4. What conclusions can you draw for process variability and process average in terms of loading time for these 5 Amazon warehouses? Which warehouse(s) need the most attention?
5. What other findings can you share with the management team?

Unloading Time:

1. What is the R bar ($\bar{R}$) value and X double bar ($\bar{\bar{X}}$) value for your sample?
2. What is the UCL and LCL of the R-Chart and X-bar Chart?
3. Plot R bar chart and X bar chart in Python.
4. What conclusions can you draw for process variability and process average in terms of unloading time for these 5 Amazon warehouses? Which warehouse(s) need the most attention?
5. What other findings can you share with the management team?