**Course Seven**

# Google Advanced Data Analytics Capstone



# Instructions

Use this PACE strategy document to record your decisions and reflections as a data professional as you work through the capstone project. As a reminder, this document is a resource guide that you can reference in the future and a space to help guide your responses and reflections posed at various points throughout the project.

# Portfolio Project Recap

Many of the goals you accomplished in your individual course portfolio projects are incorporated into the Advanced Data Analytics capstone project including:

* Create a project proposal
* Demonstrate understanding of the form and function of Python
* Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions
* Demonstrate understanding of how to organize and analyze a dataset to find the “story”
* Create a Jupyter notebook for exploratory data analysis (EDA)
* Create visualization(s) using Tableau
* Use Python to compute descriptive statistics and conduct a hypothesis test
* Build a multiple linear regression model with ANOVA testing
* Evaluate the model
* Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to solve a problem
* Articulate findings in an executive summary for external stakeholders

**Project proposal**

## **NYC Taxi Trends and Revenue Optimization Project Proposal**

***Overview***

*This project aims to analyze New York City Taxi and Limousine Commission (TLC) data to identify key factors influencing fare amounts and ride patterns. Using descriptive statistics, hypothesis testing, regression modeling, and machine learning, we will uncover actionable insights that can help increase operational efficiency and revenue for the TLC. The project will explore relationships among payment method, trip characteristics, and fare amounts and use predictive modeling to guide business strategies.*

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stages** |
| Project Planning | Define objectives, stakeholders, and research questions | Plan |
| Data Exploration | Perform data cleaning, EDA, and compute descriptive statistics | Analyze |
| |  | | --- | | Statistical Analysis |  |  | | --- | |  | | Conduct hypothesis testing on payment methods | Analyze / Construct |
| |  | | --- | | Regression Modeling |  |  | | --- | |  | | Build and evaluate a multiple linear regression model | Construct |
| |  | | --- | | Machine Learning |  |  | | --- | |  | | Build a predictive model for fare amount | Construct |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of data science**

* Who is your audience for this project?

NYC TLC decision-makers, data analysts, transportation planners.

* What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger business need?

Understand revenue trends and customer behavior to optimize pricing and payment systems.

* What questions need to be asked or answered?

Does payment method affect average fare?

What factors influence fare the most?

Can fare prediction models improve revenue forecasting?

* What resources are required to complete this project?

NYC TLC dataset, Python libraries (pandas, seaborn, statsmodels, sklearn), documentation tools.

* What are the deliverables that will need to be created over the course of this project?

Cleaned dataset, Jupyter notebook, regression model, executive summary.

**Get Started with Python**

* How can you best prepare to understand and organize the provided information?

Review data dictionary, inspect column types, handle missing values.

* What follow-along and self-review codebooks will help you perform this work?

Pandas data manipulation guide, NumPy basics, matplotlib/seaborn guides.

* What are a couple additional activities a resourceful learner would perform before starting to code?  
  Sketch EDA plan, define success metrics, scan for outliers.

**Go Beyond the Numbers: Translate Data into Insights**

* What are the data columns and variables and which ones are most relevant to your deliverable?  
  trip\_distance, trip\_duration, payment\_type, total\_amount.
* What units are your variables in?

Distance (miles), Duration (minutes), Fare (USD).

* What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?  
  Credit card users may tip more; long rides = higher fares.
* Is there any missing or incomplete data?   
  Yes, some fields contain NaN or incorrect values (e.g., zero fare).
* Are all pieces of this dataset in the same format?  
  No, some datetime or fare columns need conversion.
* Which EDA practices will be required to begin this project?

Data type conversion, outlier detection, correlation matrix, time-series plots.

**The Power of Statistics**

* What is the main purpose of this project?   
  Test whether fare amounts differ by payment method.
* What is your research question for this project?

Is there a statistically significant difference in fares between cash and credit card payments?

* What is the importance of random sampling? In this case, what is an example of sampling bias that might occur if you didn’t use random sampling?

**Importance of Random Sampling:** Ensures unbiased comparison groups.

**Bias Example:** If most long trips happen on credit cards, it could falsely inflate average fare values.

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?

NYC TLC revenue team, marketing team, operations managers.

* What are you trying to solve or accomplish?

Identify key predictors of fare amount and model their impact.

* What are your initial observations when you explore the data?  
  Strong correlation between trip distance and fare.
* What resources do you find yourself using as you complete this stage? (Make sure to include the links.)

[Statsmodels documentation](https://www.statsmodels.org), [Scikit-learn docs](https://scikit-learn.org).

* Do you have any ethical considerations in this stage?

Avoid bias against geographic regions or payment groups.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?

Predict fare amount based on trip characteristics.

* What resources do you find yourself using as you complete this stage?

Kaggle, Scikit-learn tutorials, Google Dataset Search.

* Is my data reliable?

Medium—outliers and inconsistencies present.

* Do you have any additional ethical considerations in this stage?

Ensure predictions don’t unintentionally disadvantage passengers.

* What data do I need/would I like to see in a perfect world to answer this question?  
  Include weather, traffic, time-of-day demand.
* What data do I have/can I get?  
  Trip time, distance, pickup/drop-off, payment.
* What metric should I use to evaluate success of my business objective? Why?  
  RMSE or MAE – relevant for numeric fare prediction.

**Data Project Questions & Considerations**

**PACE: Analyze Stage**

**Get Started with Python**

* Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?  
  Yes, but data must be cleaned for reliable insights.

**Go Beyond the Numbers: Translate Data into Insights**

* What steps need to be taken to perform EDA in the most effective way to achieve the project goal?  
  Clean data, calculate correlations, visualize trends, group by categories.
* Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?  
  Sort by time/date, filter invalid records, group by payment type.
* What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?  
  Bar charts for group comparisons; scatterplots for distance vs fare.

**The Power of Statistics**

* Why are descriptive statistics useful?  
  Help summarize central tendency and spread.
* What is the difference between the null hypothesis and the alternative hypothesis?  
  Null Hypothesis: Mean fares are equal for both payment types.

Alternative Hypothesis: Mean fares differ between payment types.

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?  
  Confirm linearity, check outliers, multicollinearity.
* Do you have any ethical considerations in this stage?

Avoid overfitting or misleading interpretations.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve? Does it still work? Does the plan need revising?  
  Yes, but adjustments may be needed based on data.
* Does the data break the assumptions of the model? Is that ok, or unacceptable?  
  Linearity, homoscedasticity—some violated.
* Why did you select the X variables you did?  
  Trip distance, duration, time of day.
* What are some purposes of EDA before constructing a model?  
  Feature engineering, relationship testing, normality checks.
* What has the EDA told you?
* What resources do you find yourself using as you complete this stage?  
  Coursera ML course, Scikit-learn user guide.
* Do you have any ethical considerations in this stage?  
  Ensure fairness and generalizability.

**Data Project Questions & Considerations**

**PACE: Construct Stage**

**Get Started with Python**

* Do any data variables averages look unusual?  
  Fare amounts = 0 or extremely high.
* How many vendors, organizations or groupings are included in this total data?  
  Multiple vendors and payment types.

**Go Beyond the Numbers: Translate Data into Insights**

* What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?   
  Boxplots (fare by payment), time-series (monthly trends), scatterplots.
* What processes need to be performed in order to build the necessary data visualizations?   
  Group, aggregate, visualize.
* Which variables are most applicable for the visualizations in this data project?   
  payment\_type, total\_amount, trip\_distance.
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?  
  Drop or impute with median/mean.

**The Power of Statistics**

* How did you formulate your null hypothesis and alternative hypothesis?  
  Null = no difference in mean; Alt = difference exists.
* What conclusion can be drawn from the hypothesis test?  
  Statistically significant difference between cash and credit fares.

**Regression Analysis: Simplify Complex Data Relationships**

* Do you notice anything odd?  
  Nonlinear patterns in some variables.
* Can you improve it? Is there anything you would change about the model?  
  Log-transform skewed variables, remove outliers.

**The Nuts and Bolts of Machine Learning**

* Is there a problem? Can it be fixed? If so, how?  
  Overfitting.
* Which independent variables did you choose for the model, and why?  
  Trip duration, vendor, distance.
* How well does your model fit the data? (What is my model’s validation score?)  
  R² ~0.70; room for improvement.
* Can you improve it? Is there anything you would change about the model?  
  Feature selection, regularization.
* Do you have any ethical considerations in this stage?  
  Yes—model transparency and fairness.

**Data Project Questions & Considerations**

**PACE: Execute Stage**

**Get Started with Python**

* Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing an exploratory data analysis?  
  Review zero/negative fare records.
* What data initially presents as containing anomalies?  
  Negative tips, zero durations.
* What additional types of data could strengthen this dataset?

Traffic, weather, customer ratings.

**Go Beyond the Numbers: Translate Data into Insights**

* What key insights emerged from your EDA and visualizations(s)?  
  Credit card fares are higher; long trips dominate revenue.
* What business recommendations do you propose based on the visualization(s) built?  
  Encourage card payments; promote long-distance routes.
* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?   
  Impact of time of day or borough on fares.
* How might you share these visualizations with different audiences?  
  Tableau dashboards, stakeholder presentations, and reports.

**The Power of Statistics**

* What key business insight(s) emerged from your A/B test?  
  Payment method correlates with higher revenue.
* What business recommendations do you propose based on your results?

Promote credit card usage; test fare adjustments.

**Regression Analysis: Simplify Complex Data Relationships**

* To interpret model results, why is it important to interpret the beta coefficients?  
  Distance has highest influence on fare.
* What potential recommendations would you make to your manager/company?  
  Focus on high-distance service areas.
* Do you think your model could be improved? Why or why not? How?

Yes.

* What business recommendations do you propose based on the models built?

Add categorical features like time/day.

* What key insights emerged from your model(s)?

Trip distance and duration are key drivers.

* Do you have any ethical considerations at this stage?

Avoid discriminatory pricing or regional bias.

**The Nuts and Bolts of Machine Learning**

* What key insights emerged from your model(s)?

Fare can be predicted reasonably well.

* What are the criteria for model selection?

Chose based on MAE and simplicity.

* Does my model make sense? Are my final results acceptable?

Acceptable with slight skew.

* Were there any features that were not important at all? What if you take them out?

Time of day not significant.

* Given what you know about the data and the models you were using, what other questions could you address for the team?

Can we predict tip amount? Loyalty?

* What resources do you find yourself using as you complete this stage?

UCI ML Repository, sklearn docs.

* Is my model ethical?

Yes. Transparent and avoids sensitive attributes.

* When my model makes a mistake, what is happening? How does that translate to my use case?  
  Overestimation of short trips—adjust with segment-based modeling.