Course 1: Foundations of Data Science

Background on the Automatidata scenario

Congrats on your new job as a data analyst at a data consulting firm called Automatidata. Automatidata works with its clients to transform their unused and stored data into useful solutions, such as performance dashboards, customer-facing tools, strategic business insights, and more. They specialize in identifying a client's business needs and utilizing their data to meet those business needs.

Automatidata is consulting for the New York City Taxi and Limousine Commission (TLC). New York City TLC is an agency responsible for licensing and regulating New York City's taxi cabs and for-hire vehicles. The agency has partnered with Automatidata to develop a regression model that helps estimate taxi fares before the ride, based on data that TLC has gathered.

The TLC data comes from over 200,000 taxi and limousine licensees, making approximately one million combined trips per day.

Note: *This project's dataset was created for pedagogical purposes and may not be indicative of New York City taxi cab riders' behavior.*

Project background

Automatidata is in the earliest stages of the TLC project. The following tasks are needed before the team can begin the data analysis process:

- A project proposal identifying the following:
  - Organize project tasks into milestones
  - Classify tasks using the PACE workflow
  - Identify relevant stakeholders

Team members at Automatidata and the New York City TLC

Automatidata Team Members

- Udo Bankole, Director of Data Analysis
- Deshawn Washington, Data Analysis Manager
- Luana Rodriquez, Senior Data Analyst
- Uli King, Senior Project Manager

Your teammates at Automatidata have technical experience with data analysis and data science. However, you should always be sure to keep summaries and messages to these team members concise and to the point.

New York City TLC Team Members

- Juliana Soto, Finance and Administration Department Head
- Titus Nelson, Operations Manager

Note: *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. The data shared in this project has been altered for pedagogical purposes.*

The TLC team members are program managers who oversee operations at the organization. Their roles are not highly technical, so be sure to adjust your language and explanation accordingly.

Meeting notes

Now that you are working as Automatidata's latest data analytics professional, you are given access to the company network and set up with a company email account (your first initial and last name, followed by @automatidata.org).
Opening your inbox, you notice an email from your supervisor, Deshawn.
From: Deshawn Washington
Subject: Review meeting notes
If you are able to read this, then your company accounts have been created! Now is the perfect time to get started. Last week, I attended an internal meeting with our leadership team about a new project we are about to begin. You'll receive more information in the next few days, but I would like you to be aware of some needs that were identified by our leadership team. Here is an excerpt from the notes I took during the Automatidata leadership team meeting. I've organized the points by the person who made them.
Uli King (Senior Project Manager)

- The data team will need a global-level project document to outline the goals and milestones.
- I am working closely with Titus Nelson over at the New York City Taxi and Limo Commission. He has requested some visuals to share with TLC's executives.

Luana Rodriquez (Senior Data Analyst)

- The dataset from TLC has to be inspected before any analysis can begin.
- Our team needs to determine what information the TLC data provides through exploratory data analysis (EDA).
- Eventually, our team will need to test to find if the model is delivering consistent results.

Udo Bankole (Director of Data Analysis)

- Before we present any insights to TLC, we'll need to determine whether or not the model we produce meets the project requirements.
- Once we have a final model, I'll need to know the main talking points going into our presentation with TLC.

My thoughts and concerns…

- I think it's best to use Python for the TLC project. I'll have someone on my team set that up as soon as we have the plan in place.
- It will be important to establish the relationship between any variables within the TLC data. I'd suggest the data team consider A/B testing, since that will analyze the relationship between the two most useful variables and subsequently provide data-driven support for future business decisions.

Review the meeting notes above to become familiar with the project's context. I'll ask you to identify project tasks and come up with a structure to guide the data team through this project. After our discussion about your experience in the certificate program offered by Google, I know that your efficient communication style and problem-solving will enhance the abilities of the data team.

Course 1: Foundations of Data Science

There will be more details sent to you very soon.
Welcome to the team,
Deshawn Washington
Data Analysis Manager
Automatidata
(P.S. There will be muffins in the break room every Tuesday morning. Be early…unless you like bran muffins. LOL)

Automatidata project proposal

Overview

The New York City Taxi and Limousine Commission seeks a way to utilize the data collected from the New York City area to predict the fare amount for taxi cab rides.

Milestones	Tasks	Deliverables/Reports	Relevant Stakeholder (Optional)
1	Establish structure for proj... Plan	• Global-level project document	Deshawn Washington — Data Analysis Manager
1a	Write a project proposal Plan		Uli King — Senior Project Manager

2	Compile summary informa... Analyze	• Data files ready for EDA	Luana Rodriquez — Senior Data Analyst
2a	Begin exploring the data Analyze		Deshawn Washington — Data Analysis Manager
3	Data exploration and cleani... Plan and Analyze	• EDA report • Tableau dashboard/visualizations	Luana Rodriquez — Senior Data Analyst
3a	Visualization building Construct and Analyze		Uli King — Senior Project Manager
4	Compute descriptive statis... Analyze	• Analysis of testing results between two important variables • Share results of testing	Deshawn Washington — Data Analysis Manager

4a	Conduct hypothesis testing Analyze and Construct		Udo Bankole — Director of Data Analysis
5	Build a regression model Analyze and Construct	• Review testing results • Determine the success of the model	Luana Rodriquez — Senior Data Analyst
5a	Evaluate the model Execute		Udo Bankole — Director of Data Analysis
6	Communicate final insights... Execute		
6a	Not necessary for this proj... Select PACE stage		

Note: The estimated times for the milestones in the example equate to the length of the courses where you will learn the necessary skills. Realistic timelines when working with actual clients and data scientists as a data scientist would most likely have tight deadlines, for example:

Milestone 1: 1-2 days
Milestone 2: 2-3 weeks
Milestone 3: 1 week
Milestone 4: 1 week
Milestone 5: 1-2 weeks

Project goal:
In this fictional scenario, the New York City Taxi and Limousine Commission (TLC) has approached the data consulting firm Automatidata to develop an app that enables TLC riders to estimate the taxi fares in advance of their ride.

Background:
Since 1971, TLC has been regulating and overseeing the licensing of New York City's taxi cabs, for-hire vehicles, commuter vans, and paratransit vehicles.

Scenario:
You have received notice that the recently submitted New York City TLC project proposal has been approved. The Automatidata team now has access to the New York City TLC data to analyze, identify key variables, and prepare for exploratory data analysis.

Course 2 tasks:

- Load data, explore, and extract the New York City TLC data with Python
- Use custom functions to organize the information within the New York City TLC dataset
- Build a dataframe for the New York City TLC project
- Create an executive summary for Automatidata

Note: *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

2017_Yellow_Taxi_Trip_Data.CSV

408,294 rows – each row represents a different trip
18 columns

Column name	Description
ID	Trip identification number

VendorID	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged
RateCodeID	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before being sent to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip

Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.