Parcel Perform

Data Analyst Position

Take home exercise

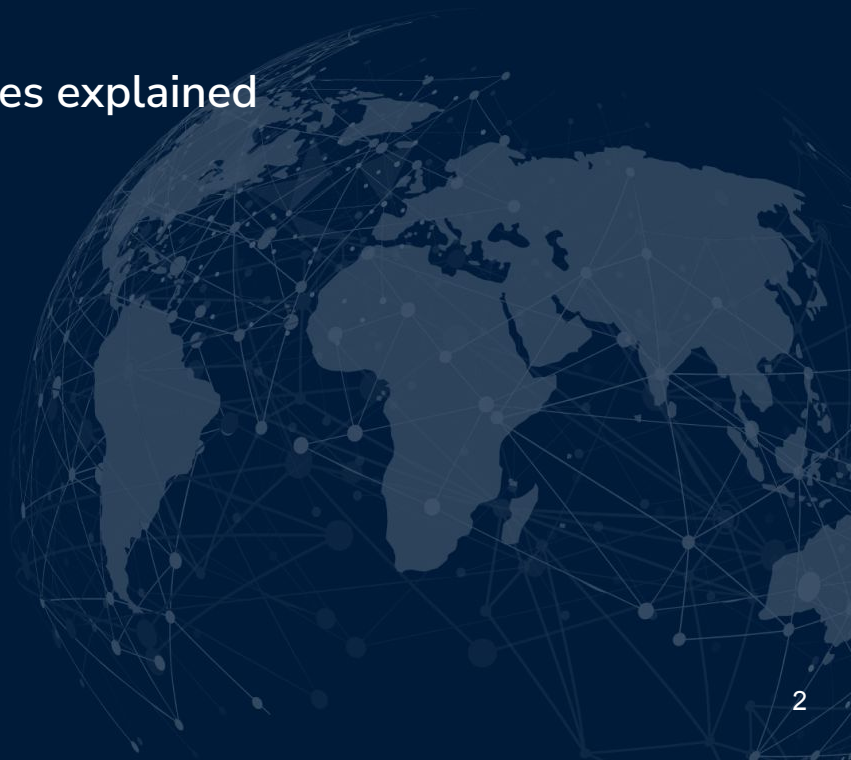
2025

parcelperform

Content

- Instruction
- Terminologies explained
- Part 1
- Part 2
- Part 3

  (FAQ)



parcelperform

Instruction

- For this take-home exercise, you will be given 3 separated exercises. The detailed instructions will be written in each exercise. You can consolidate your answer in one file or in multiple files, as long as the title of each file is clear for us to review (example: part 1 submission, part 2 submission,...)
- Feel free to use any appropriate tools, techniques, and methodologies to extract insights and draw meaningful conclusions. However, you need to **document the steps and techniques** you used during the analysis process. This should cover data cleaning, transformation, analysis methods, and any assumptions made.
- The time we estimate to finish the exercise is 8 hours. We give you 4 days to complete it. Should you have any questions, feel free to contact our team via email, your concern would be answered right away.

parcelperform

Terminologies explained

- **Carrier**: The logistics carrier delivering the parcel, e.g. DHL, Ninja Van
- **Origin country**: Country where the parcel is sent.
- **Destination country**: Country where the parcel is received.
- **Pick-up**: Time and date when the carrier collects (= picks up) parcel from sender.
- **Out for Delivery**: Timestamp when the carrier starts to deliver the parcel to recipient.
- **First attempt**: Timestamp when the carrier for the first time attempts to deliver a parcel.
- **Final delivery**: Timestamp when the recipient receives their parcel.
- **Transit time**: The time between when parcel is picked up and and when it delivered to recipient.
- **Domestics Trade lane**: Tradelane where origin and destination country is the same
- **Estimated Delivery Date (EDD)**: The projected date when a customer can expect to receive their order

Additional Params explained (for question 2 and 3)

- **old_parcel_expected_time_first_start / old_parcel_expected_time_latest_start**
  - The start time of the current EDD we receive from the carrier.
- **old_parcel_expected_time_first_end / old_parcel_expected_time_latest_end**
  - The end time of the current EDD we receive from the carrier.
- **new_parcel_expected_time_first_start / new_parcel_expected_time_latest_start**
  - The NEW start time of the EDD we receive from the carrier.
- **new_parcel_expected_time_first_end / new_parcel_expected_time_latest_end**
  - The NEW end time of the EDD we receive from the carrier.

Note: You can freely choose which approaches are most suitable for indicating the EDD (choose old value, new value, start point, end point, or average of two values,...). Please state the reason for your approach in the submission.

Part 1

You are given two tables, with the corresponding schema as below:

PARCEL TABLE	
parcel_id	int
carrier_name	varchar(500)
pick_up_date	timestamp
out_for_delivery_date	timestamp
first_attempt_date	timestamp
final_delivery_date	timestamp
origin_country	varchar(500)
destination_country	varchar(500)
is_delivered	boolean

LOG TABLE	
log_id	int
parcel_id	int
raw_log_description	varchar(500)
log_key	varchar(500)
log_timestamp	timestamp
additional_params	varchar(500)

Part 1

Questions

Using SQL language, please provide answers to the following questions:

1. How to find out is the average, median and 90th percentile transit time (in days) of all domestics trade lanes in 2024?
2. What is the max transit time (in days) and how many parcels has this transit time?
3. Pick the top 2 carriers in terms of volume that is operating in each trade lane.
4. How to find out parcels that is delivered but has no record in log table?
5. Find rows with the same parcel_id but with different carrier name, concatenate all carriers into a new result column named 'list_of_carrier'.

   Example output:
   - parcel_id | list_of_carrier
   - ----------| ----------------------
   - 1234567 | Ninja Van; DHL; UPS;

Note:

- Use the schema and provide correct SQL syntax is expected. The data files are used for questions 2&3. Hence, you do not need to extract answers in concrete numbers from the data files.

parcelperform

Part 2

Questions

You are given two separated data files in parquet format, namedly parcel_table and log_table. Feel free to use these data files to explore the data and present your findings.

We want you to setup an report on the estimated delivery date (EDD) accuracy. Which carrier has the highest accuracy? Is there any pattern with the provided estimated delivery date?

Note:

- Use Python language to perform this question. (libraries such as Polars, Pandas... are allowed)
- Your finding must be put into some of the ready-to-present documentation as part of the submission (Can be iPython, Powerpoint, MD file, html,...). Please attach the work file for reference too (your script in python, Jupyter notebook,...).

parcelperform

Part 3

Questions

You still work on the two data files provided in part 2. This time, we want you to further explore the data and present your findings. The finding can be anything from parcel behavior, to operation suggestion.

- If you're not sure where to start with, there are some initial questions you can look at:
  - How many parcels completed its delivery journey?
  - Is parcel always delivered on its first attempt?
  - What is the delta (change) in estimated delivery date for each carrier?
  - What is the coverage of estimated delivery date?
  - ...
- **Note**:
  - Use Python language to perform this question. (libraries such as Polars, Pandas,... are allowed)
  - Your finding must be put into some of the ready-to-present documentation as part of the submission (Can be iPython, Powerpoint, MD file, html,...). Please attach the work file for reference too (your script in python, Jupyter notebook,...).

FAQ:

1. **Ques**: What does "90th percentile transit time" mean?

   Ans: Percentile is a statistical concept used to describe the relative position of a value within a dataset. It indicates the percentage of data points that fall below a given value. For example, the 90th percentile means that 90% of the data points are below that value, and 10% are above. Percentiles are widely used in data analysis for identifying outliers, customer segmentation, performance benchmarking, and risk analysis. You can explore the SQL syntax about percentiles for more information.

2. **Ques**: As I understand, "transit time" is the time between "Out for Delivery" and "Pick-up". Is this correct?

   Ans: No, as explained on page 4 under "Terminologies Explained," "transit time" refers to the duration between when a parcel is picked up and when it is delivered to the recipient. Therefore, it must be based on the Pick-up and Final Delivery times.

3. **Ques**: What is "delta"?

   Ans: In general, "delta" refers to the change or difference between two values. It can be represented by the symbol Δ (delta), often pronounced as "delta." For example, if you have two data points, A and B, the delta value would be the numerical difference between them, calculated as B - A.

   In the context of this exercise, delta value change is the change of value from old to new parcel_expected_time.

FAQ:

4. **Ques**: How can I find the "estimated delivery date" (EDD) from the data provided?

   Ans: You can find it in the additional_params column. I will explain it shortly to you here:

- "first" vs. "latest": They are technically the same, so no need to differentiate. Clean column names if needed.
- "start" vs. "end": If the carrier provides EDD as a range (e.g., 2023-06-08T00:00:00+00:00 to 2023-06-09T23:59:59+00:00), we record:
  - new_parcel_expected_time_latest_start = 2023-06-08T00:00:00+00:00
  - new_parcel_expected_time_latest_end = 2023-06-09T23:59:59+00:00
- "old" vs. "new": Tracks updates to delivery estimates. If the carrier updates the estimate (e.g., from 2023-06-06T14:00:00Z to 2023-06-09T00:00:00Z), we record:
  - old_parcel_expected_time_first_start = 2023-06-06T14:00:00Z
  - new_parcel_expected_time_first_start = 2023-06-09T00:00:00Z

The end

parcelperform