Cristopher,

Thank you for taking the time to meet with me yesterday and apologies for all the noise during the interview.

Please see below for the technical challenge.

Please complete this by Thursday May 26th at 11:59P EST.

Please reply to this email with any questions you may have about the challenge.

**Dataset and Context**

[**Bus Breakdown and Delays | NYC Open Data**](https://data.cityofnewyork.us/Transportation/Bus-Breakdown-and-Delays/ez4e-fazm)

* The Bus Breakdown and Delay system collects information from school bus vendors operating out in the field in real time.
* Bus staff that encounter delays during the route are instructed to radio the dispatcher at the bus vendor’s central office.
* The bus vendor staff are then instructed to log into the Bus Breakdown and Delay system to record the event and notify OPT.
* OPT customer service agents use this system to inform parents who call with questions regarding bus service.
* The Bus Breakdown and Delay system is publicly accessible and contains real time updates.
* All information in the system is entered by school bus vendor staff.

**Task**

1. **Write working code (in python only) to demonstrate an** [**ETL workflow**](https://en.wikipedia.org/wiki/Extract,_transform,_load) **on this dataset.**
   1. *E = Code that demonstrates how you would extract this data from the source.*
   2. *T = Code to demonstrates simple transformations on the data you’ve extracted. Transformations can include data cleaning or data-reshaping to support your analysis*
   3. *L = Code that loads this data to supports your Insights.*
2. **Report insights on a Google Slides deck containing Insights on:**
   1. Data quality (missing data)
   2. Temporal insights (Insights into Breakdowns by Time)
   3. Geographical insights
   4. Other insights that support decision-making.

**Deliverables ( due by Thursday May 26th at 11:59P EST)** *Please reply to this email with 2 deliverables.*

1. Working Python Code hosted on<https://colab.research.google.com/>
   1. Here’s an introduction on how to use Colab <https://colab.research.google.com/notebooks/intro.ipynb>
2. A slide deck on Google Slides containing Insights

**Evaluation /** Candidate submissions shall be evaluated on

1. Functioning of Code
2. Approaches used to perform the ETL task
3. Quality of their Insights

Thanks!

DATA Note:

Dataframe initial Schema:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 493067 entries, 0 to 493066

Data columns (total 21 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 School\_Year 493067 non-null object

1 Busbreakdown\_ID 493067 non-null int64

2 Run\_Type 493063 non-null object

3 Bus\_No 493055 non-null object

4 Route\_Number 493060 non-null object

5 Reason 493065 non-null object

6 Schools\_Serviced 493060 non-null object

7 Occurred\_On 493067 non-null object

8 Created\_On 493067 non-null object

9 Boro 481863 non-null object

10 Bus\_Company\_Name 493067 non-null object

11 How\_Long\_Delayed 441256 non-null object

12 Number\_Of\_Students\_On\_The\_Bus 493067 non-null int64

13 Has\_Contractor\_Notified\_Schools 493067 non-null object

14 Has\_Contractor\_Notified\_Parents 493067 non-null object

15 Have\_You\_Alerted\_OPT 493067 non-null object

16 Informed\_On 493067 non-null object

17 Incident\_Number 10836 non-null object

18 Last\_Updated\_On 493067 non-null object

19 Breakdown\_or\_Running\_Late 493067 non-null object

20 School\_Age\_or\_PreK 493067 non-null object

dtypes: int64(2), object(19)

First observations:

1. 98 % of incident Number values are null