Summary:
The delivery duration can be predicted by this model on average about 10 minutes error, for example, if the model predict one delivery is 50 minutes, then the actual delivery time will be 40-60 minutes. We can use this model to estimate the delivery, and if the predict delivery time is really long, we can take some actions before that, and make customer happier.

Data Process:

- Dependent Variable:
    1. Drop 7 records with missing deliver time.
    2. Create dependent variable "length" as defined.
    3. Safely drop 3 outliers (bigger then 60000), it is not good to drop outliers, because sometimes the outlier is the thing needed to be analysis, but this one is really big and bad.
    4. Log-transform the length, because the length has a long tail, it is good to log transform and predict it.
- Independent Variables:
    1. Safely drop 3 outliers.
    2. Create new features, like hour, weekday, deliver hour, deliver weekday, average price, estimated total duration.
    3. Replace some negative value with missing value, and fill in to other values later (min item price, total_onshift_dashers, total_busy_dashers, total_outstanding_orders.
    4. Missing Filling:
        - store_primary_category: using most common category in each store id group, if the missing is the most common, then fill in 'OTHER'.
        - market_id : using most common category in each store id group, if the missing is the most common, then fill in '-1'.
        - order_protocol: same logic as market id.
        - estimated_store_to_consumer_driving_duration: the missing ratio is low, just use the median of the whole data.
        - min_item_price: replace negative value into median for each store id.
        - total_outstanding_orders is follow patterns by hours and market id, fill in median within these two groups.
    5. Convert some Category values to string: market id, order_protocol, hour.
    6. total_busy_dashers&total_onshift_dashers is not good in the data, about 30% is either missing or "busy driver #" is bigger than "total dashers #". I imputed these bad values using median in each hour and market id group. And using model to compare "drop these fields" and the model "impute values", and the model with impute values fields is better.
    7. Add log-transformed to some tailed independent variables.
    8. Tried to add squared variables to some numeric predictors, but not improving model.
    9. Drop some unusable variables to create the final data.

- Draw correlation heat map, since we don't have too much features, I include most of the usable features, and tree-based model is not sensible to less useful features, and I use LASSO and ridge to give penalty to less useful features.
- Four models are used, LASSO, Ridge, XGboost, GBR. I only use gridsearchcv on ridge model, because time is limited.
- Stack the two best performance models xgboost and gbr, give the equal rate of 0.5, and then using lasso to model the stack.
- XGBoost model is better than stack, if we got more time to improve GBR, maybe we can get better stack model. Now using XGBoost model.
- I use 30% of the original data as a testing data and get the RMSE score for the log of length to evaluate the final result.
- Since the time is limited and information is limited, I can only try some of the parameters of the models and some of the feature transformation.
- Suggestion:

1. May need to discuss with data team or CS team to get accurate available dasher and busy dasher data.
2. There is a spike on 8 am, we may need to add dashers in the morning, and I am not sure why we don't have business from 9 to 13, maybe a lot of traffic? We may give some promotion for the dasher during the morning to let more people involved.
3. When the outstanding orders increased, the length of the delivery duration increase dramatically. That make sense, because of the busy time. Maybe we can send some notification to let more dasher come when the outstanding order is big.