By: Adam Brenner, Benjamin Dinhofer, Jai Ramotar

The Problem


Nous cherchons un chauffeur pour vous
Restez appuyé pour annuler


When I say
ETA,
I mean this.
ARRIVAL
TIME
Estimated Time
Of Arrival

Original Data

- Target is ETA.
- The only readily useful column is Trip_distance.

	ID	Timestamp	Origin_lat	Origin_lon	Destination_lat	Destination_lon	Trip_distance	ETA
18266	7SGG91P5	2019-11-21T23:49:20Z	3.037	36.729	2.995	36.737	8477	751
78477	XMLV1OI4	2019-12-03T23:12:40Z	3.060	36.780	3.056	36.773	2640	384
24656	AK1M1PJD	2019-12-13T21:22:16Z	3.020	36.753	3.207	36.698	21405	1083
27982	BZNGXBRU	2019-12-13T21:34:49Z	3.015	36.754	3.074	36.752	12864	921
34960	F1T9FUWX	2019-11-27T20:27:10Z	3.175	36.737	3.163	36.715	4661	771

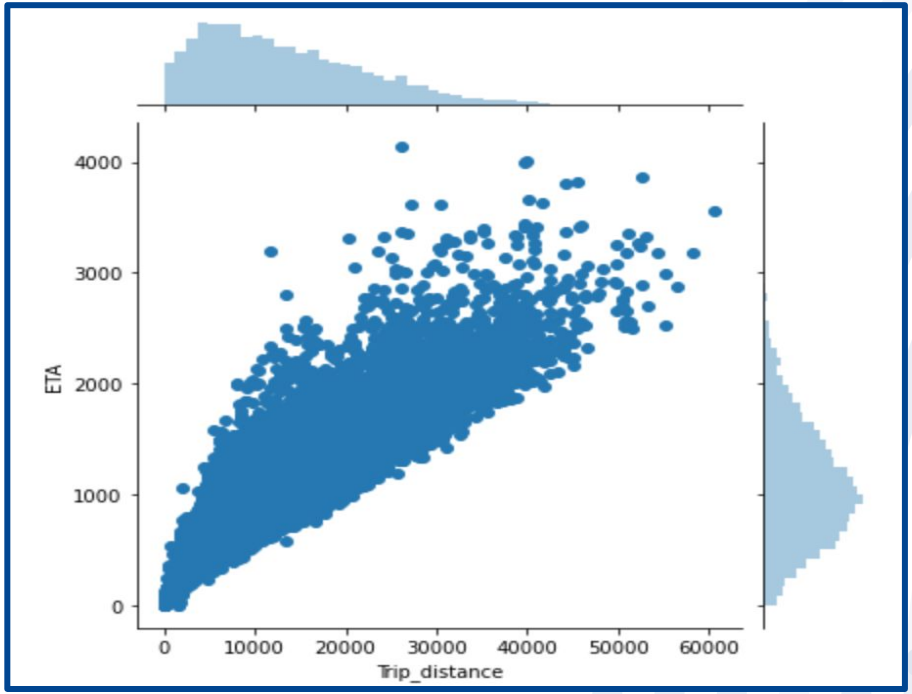
Feature Engineering: Datetime

- Expand on Timestamp Column by converting Timestamps to Datetime objects.
- New Columns:
  - Hour = hour of day
  - Weekday = day of week
  - Is weekend = if day is saturday or sunday

hour	weekday	is_weekend
23	3	No
23	1	No
21	4	No
21	4	No
20	2	No

EDA: Trip_distance

- Trip Distance was the most heavily correlated with ETA.



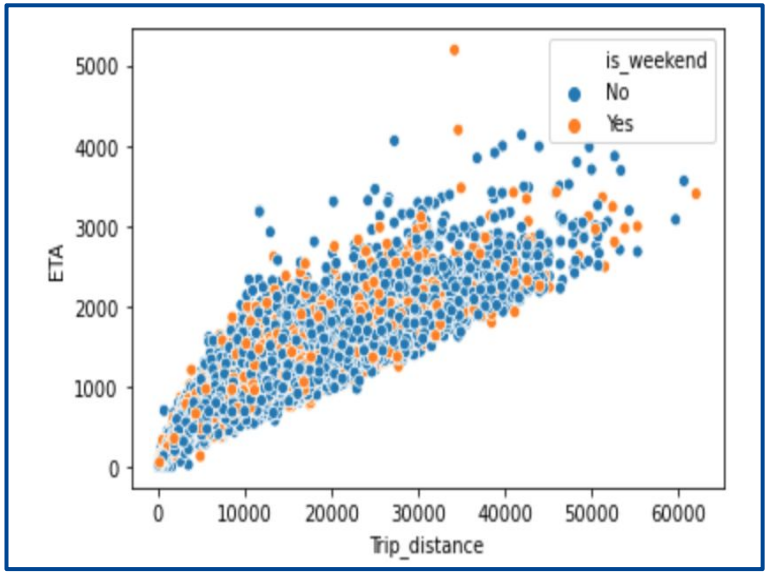
Feature Engineering: Encoding

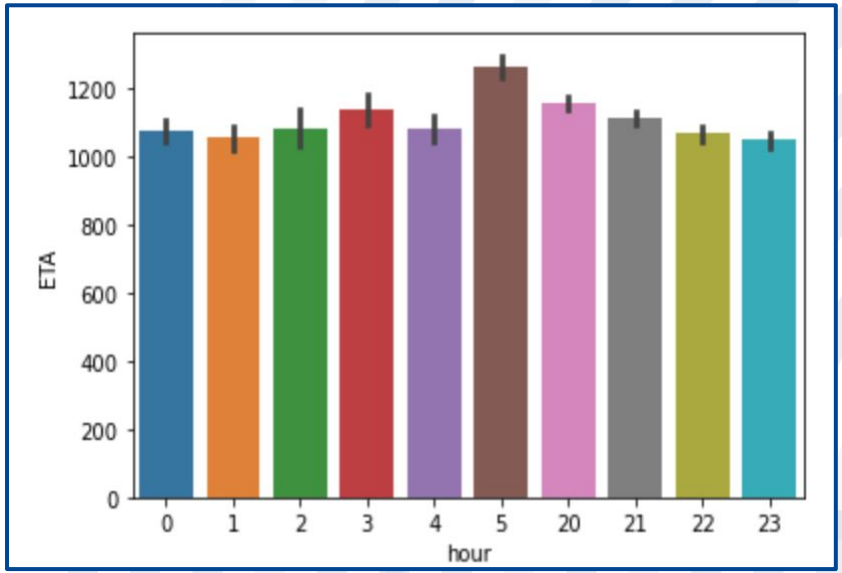
- Can drop “weekday” column because of its redundancy when used with “is_weekend” column.

Timestamp	hour_1	hour_2	hour_3	hour_4	hour_5	hour_20	hour_21	hour_22	hour_23	is_weekend_Yes
2019-11-21 23:49:20+00:00	0	0	0	0	0	0	0	0	1	0
2019-12-03 23:12:40+00:00	0	0	0	0	0	0	0	0	1	0
2019-12-13 21:22:16+00:00	0	0	0	0	0	0	1	0	0	0
2019-12-13 21:34:49+00:00	0	0	0	0	0	0	1	0	0	0
2019-11-27 20:27:10+00:00	0	0	0	0	0	1	0	0	0	0

EDA: Weekends and Time of Day

- Neither Weekend or hour seem to have a significant impact on ETA.





Model 1: Ridge Regression

- First model used **ridge regression.**
- Was evaluated using **cross validation.**
- Received an **r-squared** test score of **0.814.**

	fit_time	score_time	test_score	train_score
ridge_reg	0.012488	0.003058	0.814453	0.81613

Can We Improve This Model?

Outside of distance, our features didn't have much effect on ETA, so let's start by adding new features.

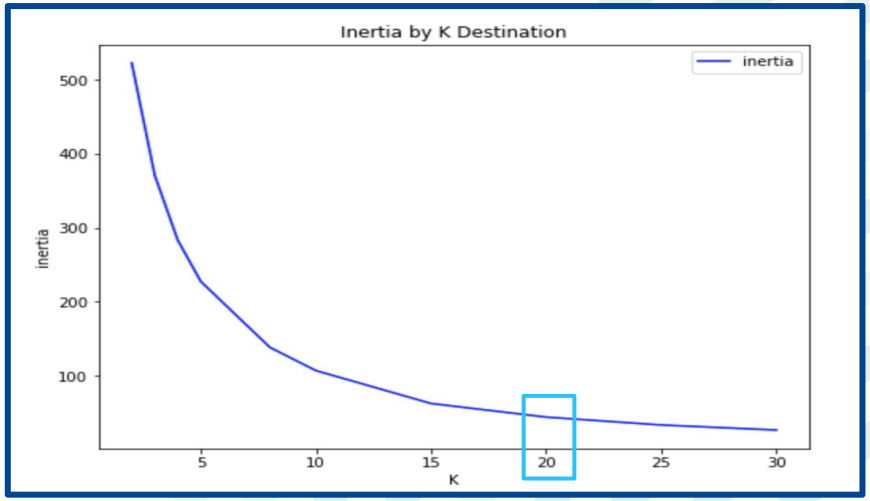
Feature Engineering: Location

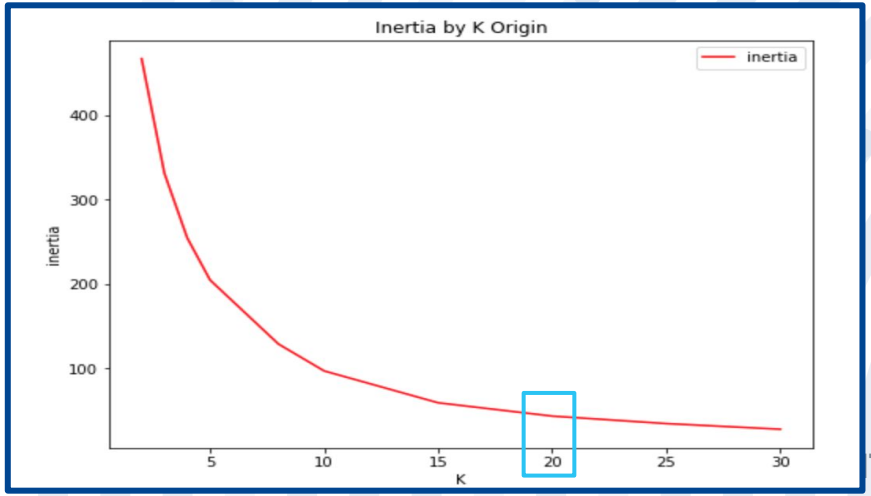
- The Original Dataset gives raw geographical information in the latitude and longitude.
- Can build features based on this data to improve accuracy.

	ID	Timestamp	Origin_lat	Origin_lon	Destination_lat	Destination_lon	rip_distance	ETA
18266	7SGG91P5	2019-11-21T23:49:20Z	3.037	36.729	2.995	36.737	8477	751
78477	XMLV1OI4	2019-12-03T23:12:40Z	3.060	36.780	3.056	36.773	2640	384
24656	AK1M1PJD	2019-12-13T21:22:16Z	3.020	36.753	3.207	36.698	21405	1083
27982	BZNGXBRU	2019-12-13T21:34:49Z	3.015	36.754	3.074	36.752	12864	921
34960	F1T9FUWX	2019-11-27T20:27:10Z	3.175	36.737	3.163	36.715	4661	771

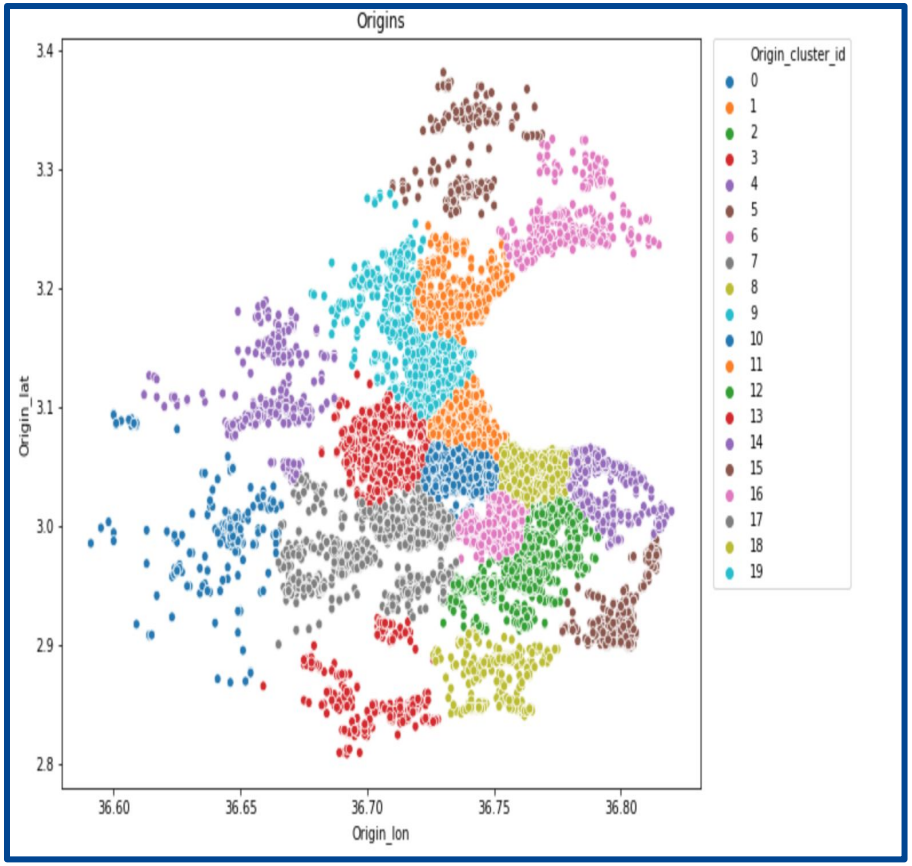
K-means Clustering

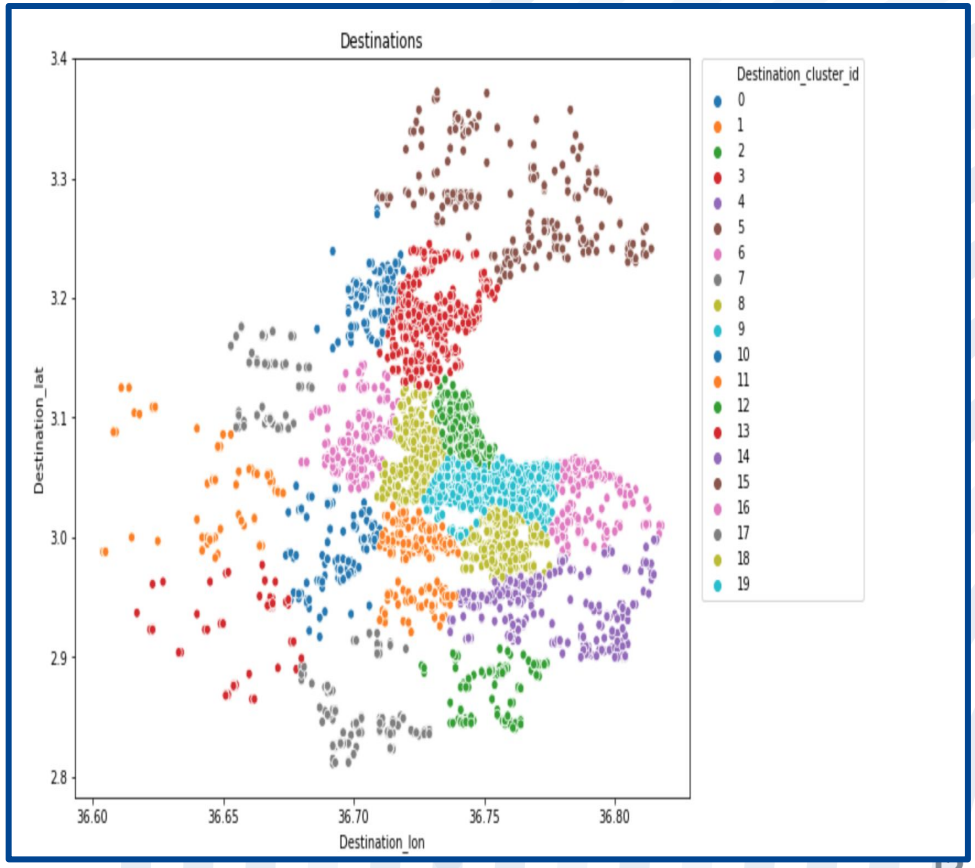
- **K-means clustering** allowed us to group the location data based on their latitude and longitude
- The **Elbow method** revealed that the optimal number of clusters was 20.
- Received silhouette score of around .45 for both groups.
- Intuitively, this model assigned each origin point and each destination point to a region.





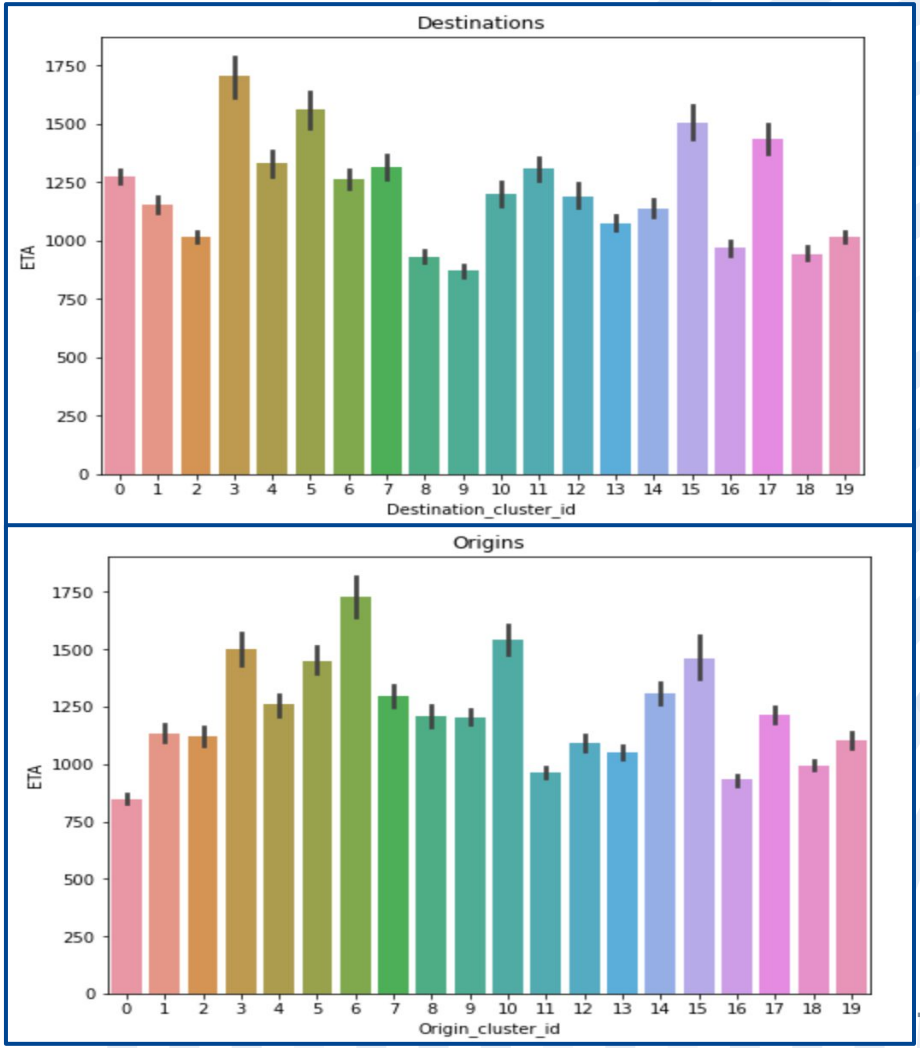
K-Means Visualized





Cluster ETA

- These graphs show that each different **region** has a **significant effect** on the **ETA** of the ride for both origin and destination.
- We **one-hot-encoded** the cluster id's for the origin points and destination points in order to implement this into our model.



Hyperparameter Tuning

- In addition to new features, **hyperparameter tuning** would increase the accuracy of our model.
- We used the **randomized search** method and decided to focus on the **alpha** and **fit-intercept** parameters.

Results

```
random_search_ridge.best_score_
```

```
0.8435887133016168
```

```
random_search_ridge.best_estimator_
```

```
Ridge(alpha=0.18204445369091374)
```

Model 2: Ridge Regression with Locations

- Adding the new features and hyperparameter tuning, we fit our model again with **ridge regression**.
- Our new results after **cross validation**:

	fit_time	score_time	test_score	train_score
Random_Search_Ridge	2.90832	0.003134	0.845643	0.853359

- Our **r-squared** score increased from **0.814** to **0.846**.
- An overall increase of about **3%**

Let’s try using an ensemble model.

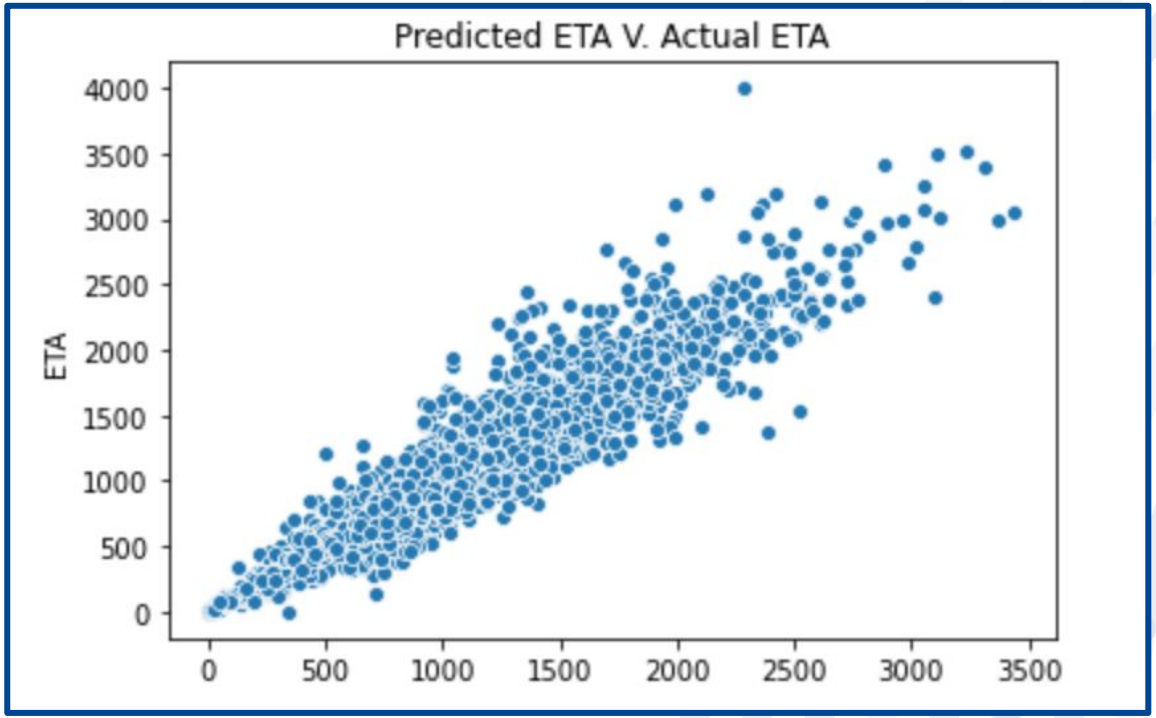
Random Forest Model

- To improve upon the model even further, we decided to use the **random forest regression** model.

	fit_time	score_time	test_score	train_score
Random_Search_Ridge	3.503533	0.005829	0.845732	0.853347
randomforest_100	1.444659	0.037692	0.850453	0.978334

- We left out hyperparameter tuning because it made the model take too long to run.
- Our **r-squared** score increased again, but less significantly.
- There was an increase of about **1%.**

Predictions



Insights

- The most important regressor was distance.
- Time of day and if the day was a weekend had little effect on the model.
- Using the geographical along with hyperparameter tuning improved our model.
- Geographical data was more useful than time related data.
- The most effective model was the random forest regression model that included the geographical data.

Improvements

Thanks!

Any questions?