Predicting Offline Conversions

— Eugenie Chen

Agenda

- Background
- Data
- Exploratory Analysis
- Modeling and Evaluation
- Next Steps

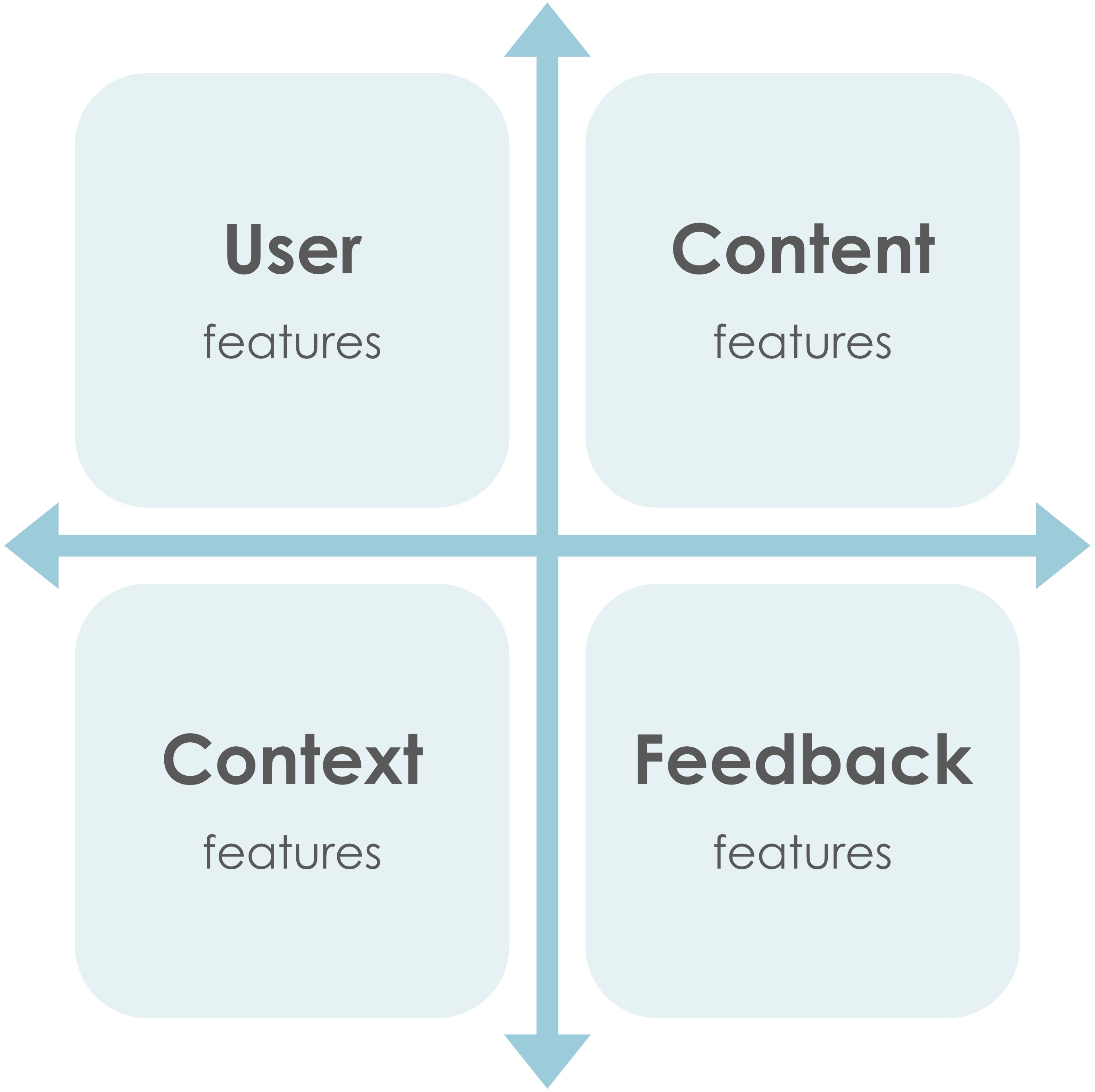
Background

- Trends seen in online media
  - Demand for performance dependent pricing model
  - Accurate response prediction is key to maximize efficiency and revenue
- Challenges and opportunities for offline media
  - Lack of attribution solutions for out-of-home media
  - RTB (real-time bidding) nearly impossible for programmatic out-of-home buyers
  - Opportunity to optimize transaction models as well as to inform media planning

Data

- Age
- Gender
- Education
- Employment
- …

- Venue type
- Date & time
- Targeting strategy
- Screen location
- …



- creative_id
- Formats
  - video vs static
- …

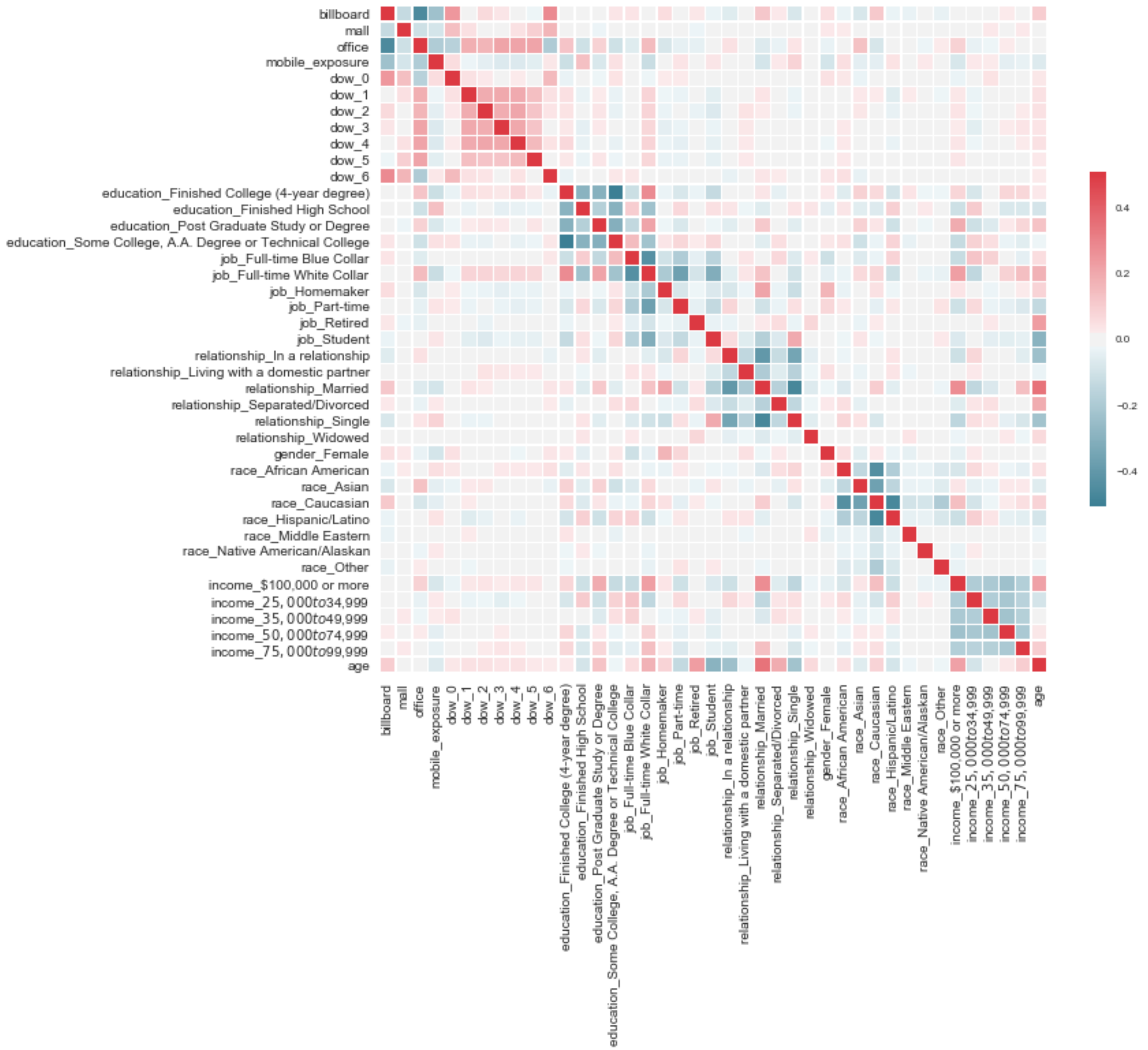
→ Not available today
- for next steps

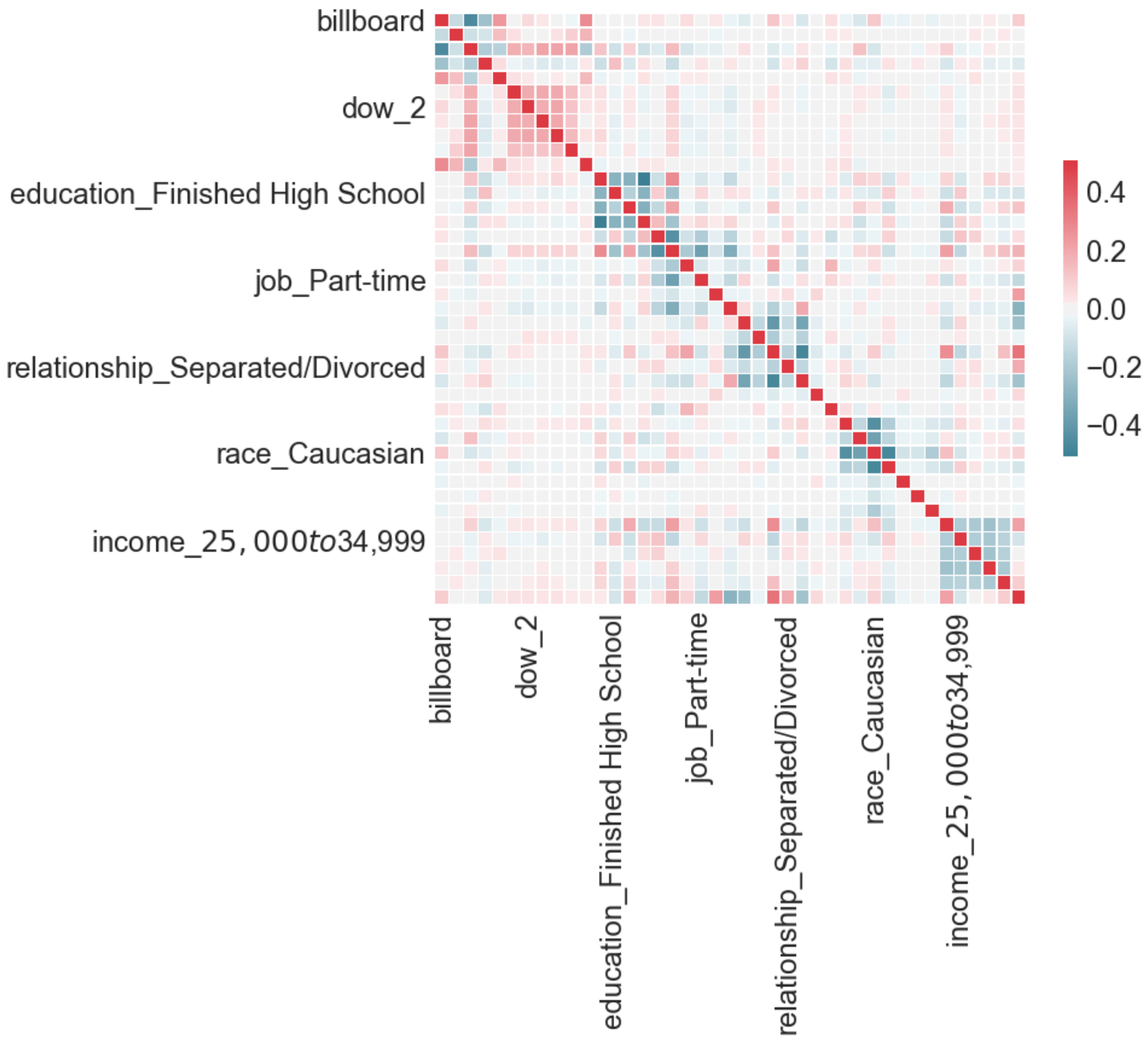
SQL – pulling distance feature

```
--Step 1: add geometry columns to tables


alter table exposure_table
add column geom geometry(point,4326);

update exposure_table
set geom = ST_SetSRID(ST_MakePoint(longitude::float,
latitude::float),4326);


alter table conversion_table
add column geom geometry(point,4326);

update conversion_table
set geom = ST_SetSRID(ST_MakePoint(longitude::float,
latitude::float),4326);


--Step 2: calculate distance between point of exposure
and nearest store location at time of exposure
```

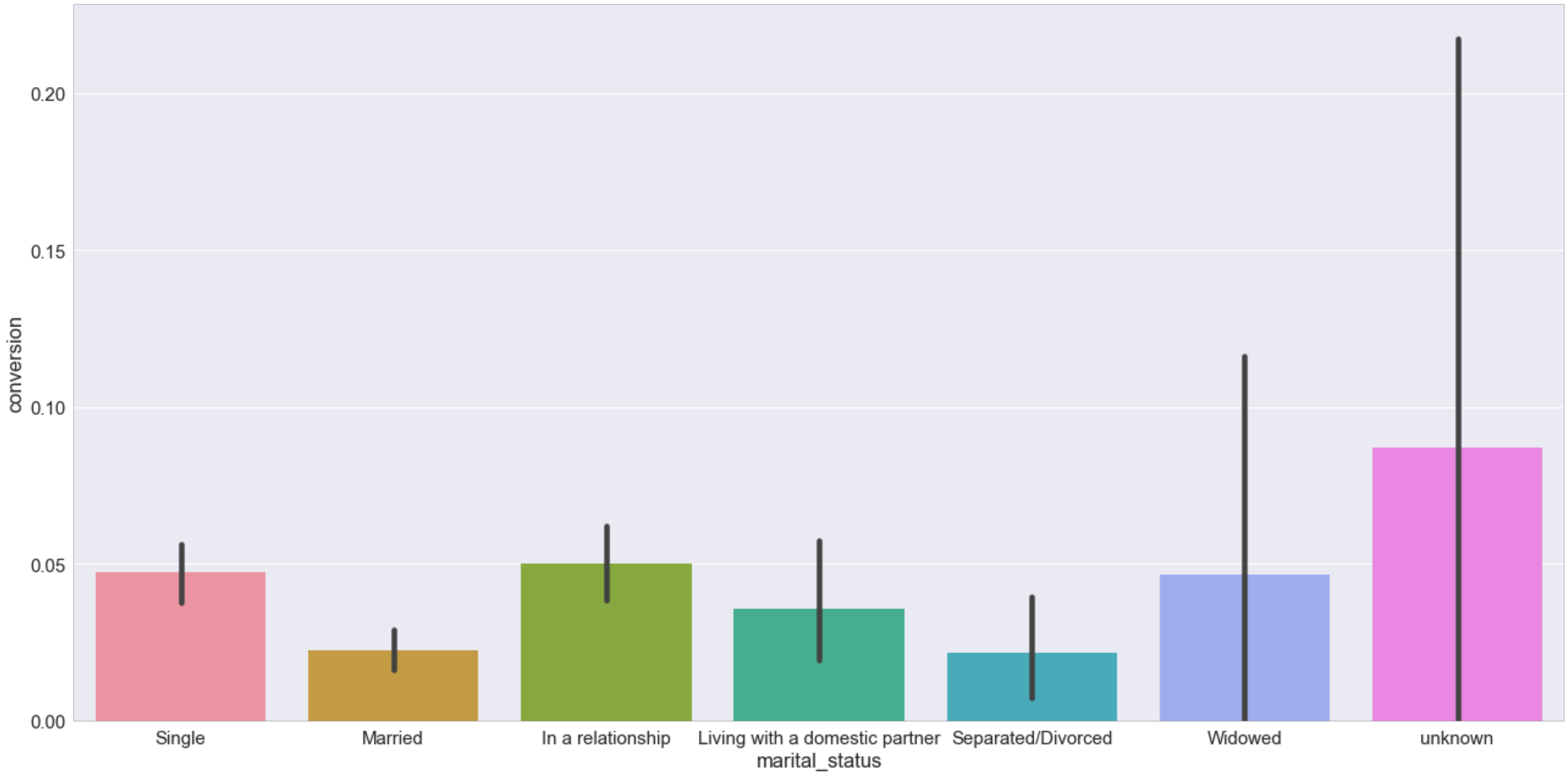
```
SELECT
    a.user_id,
          a.latitude as exp_lon,
          a.longitude as exp_lat,
          b.store_id,
          b.latitude as store_lat,
          b.longitude as store_lon,
          ST_Distance(a.geom, b.geom) as dist
INTO exp_store_dist
FROM
    exposure_table a,
          conversion_table b
  ORDER BY
    a.geom <->
    b.geom;

SELECT
          a.*
FROM exp_store_dist a
JOIN
          (SELECT
                    DISTINCT user_id,
                    min(dist::float) AS dist
          FROM exp_store_dist
GROUP BY 1) b
ON a.user_id = b.user_id
AND a.dist = b.dist;
```

Exploratory Analysis – collinearity

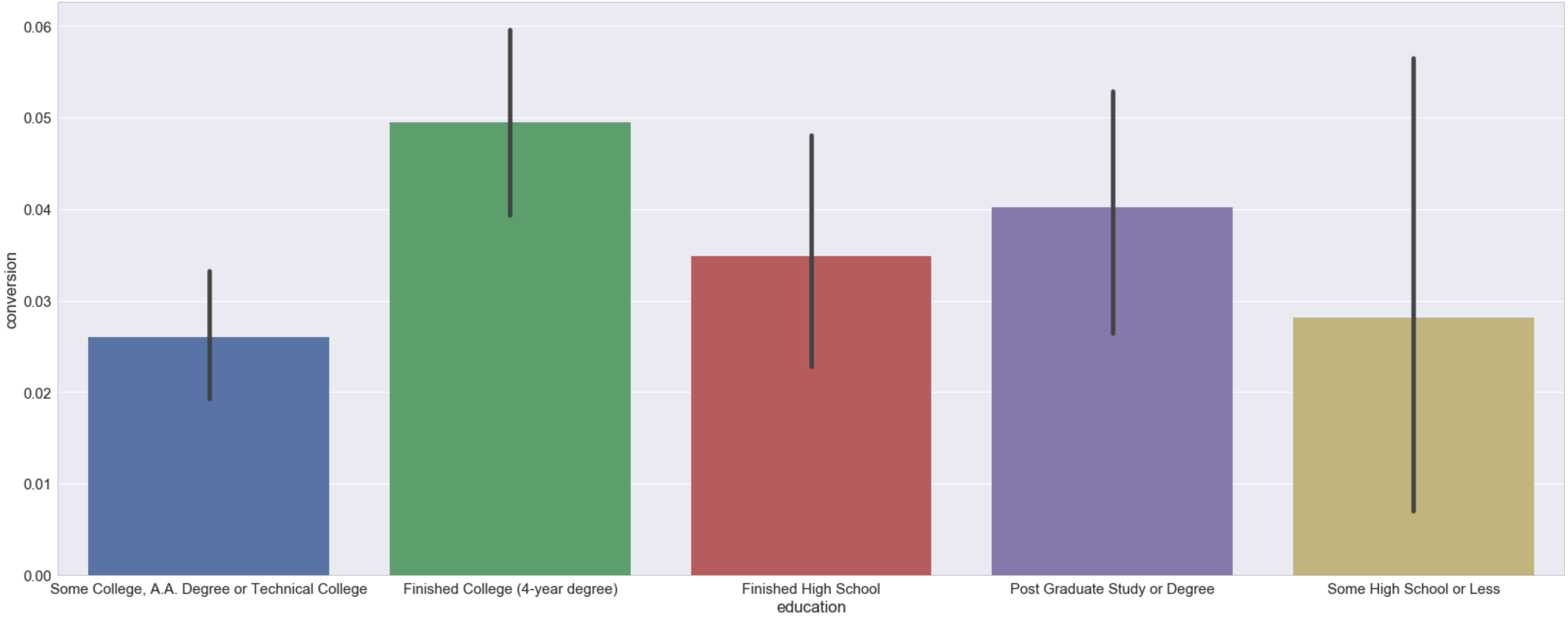




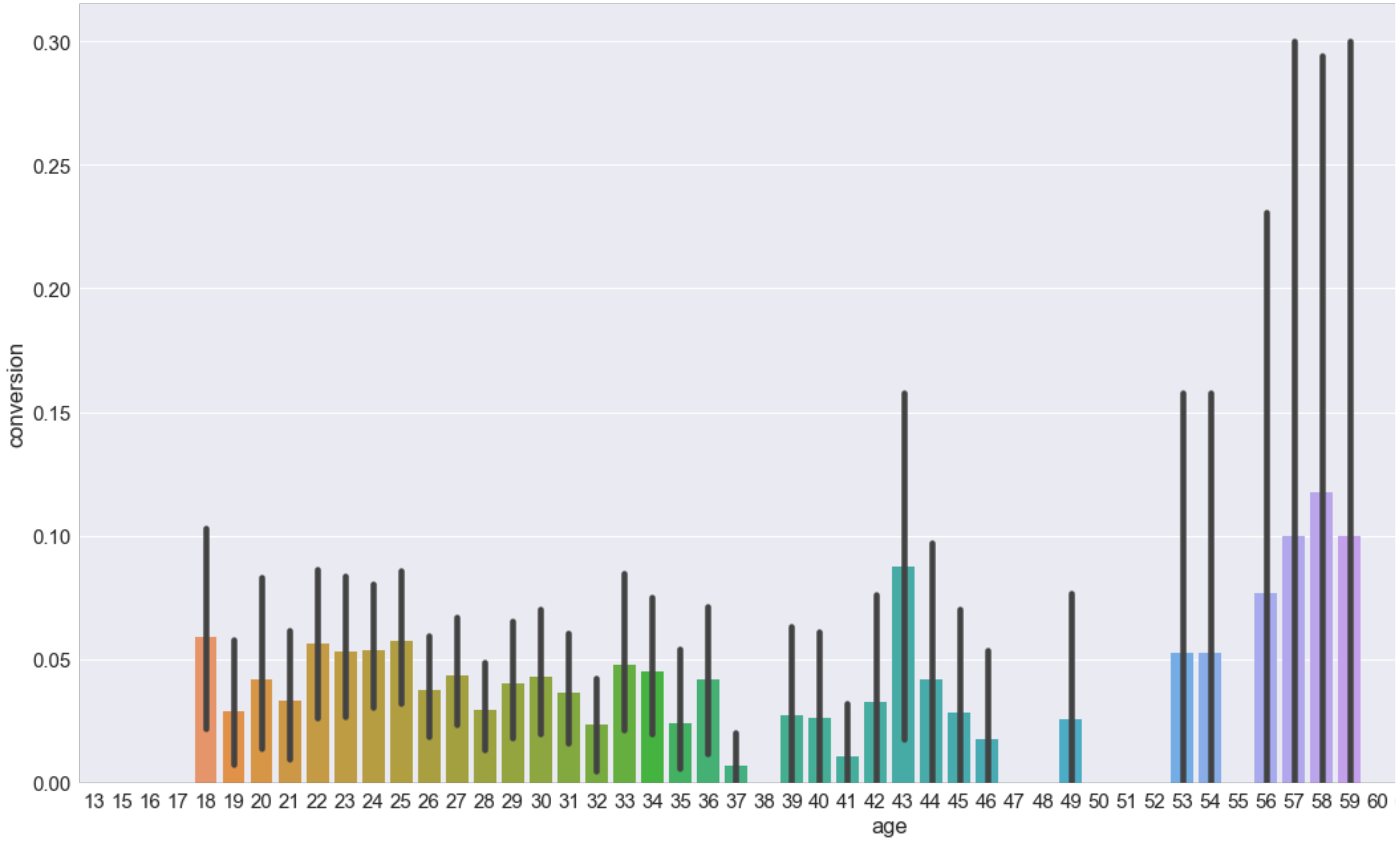
Exploratory Analysis – marital status vs. conversion

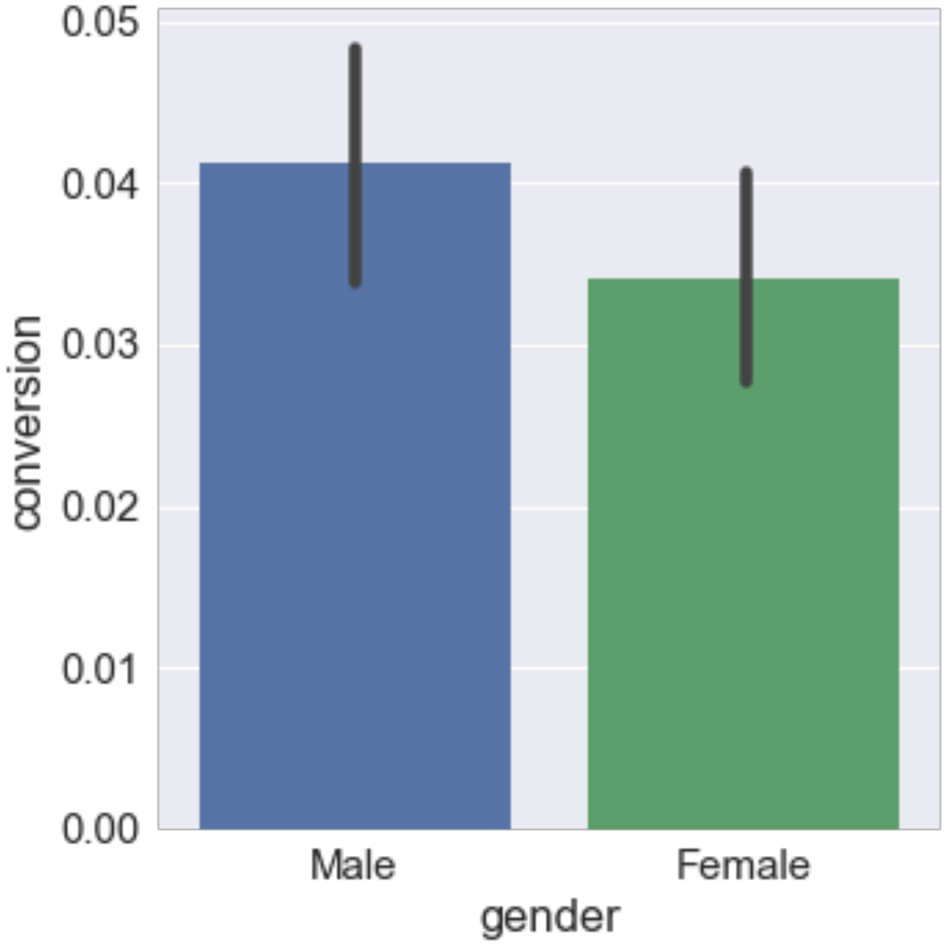


Exploratory Analysis – education vs. conversion

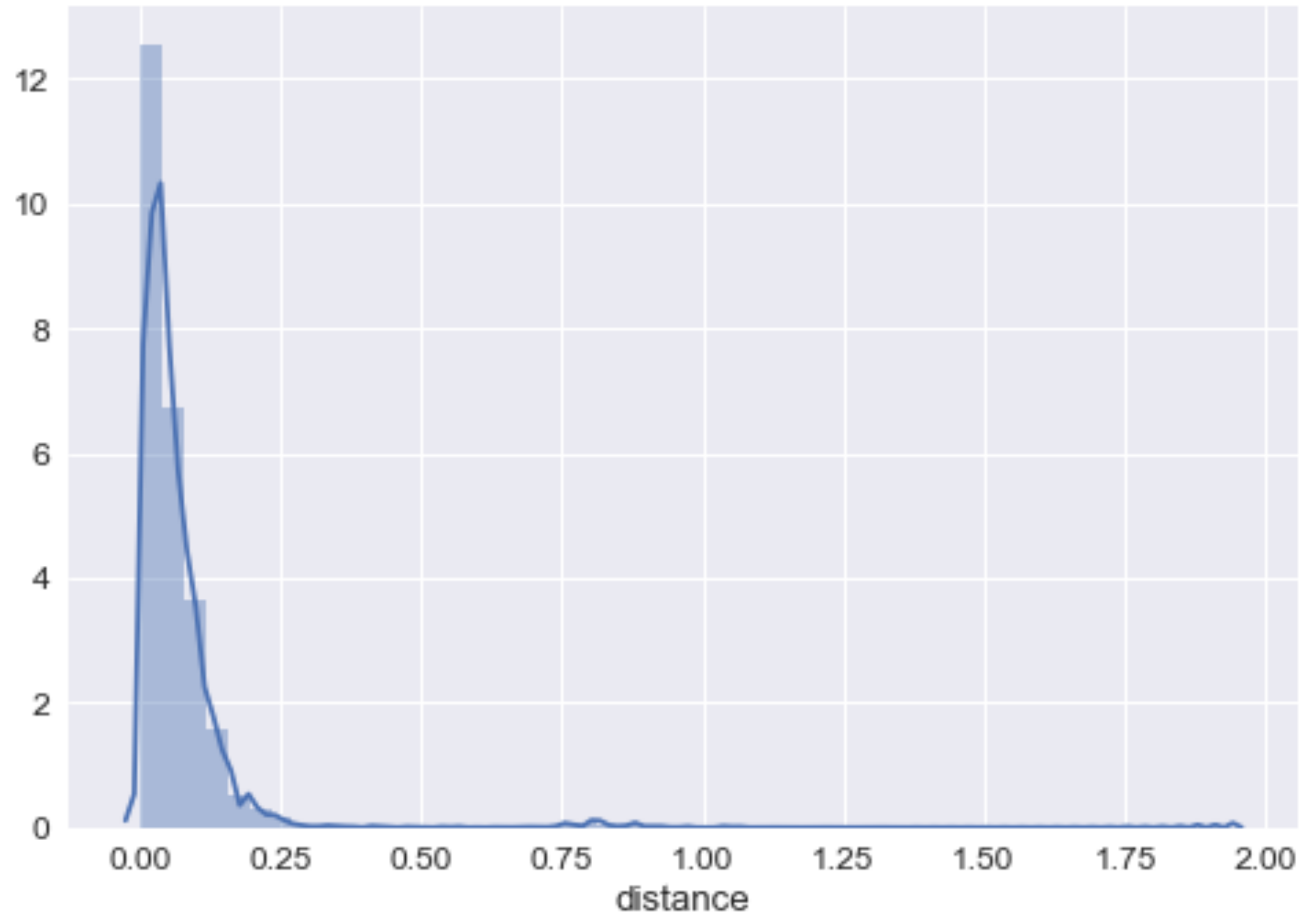


Exploratory Analysis – age & gender vs. conversion

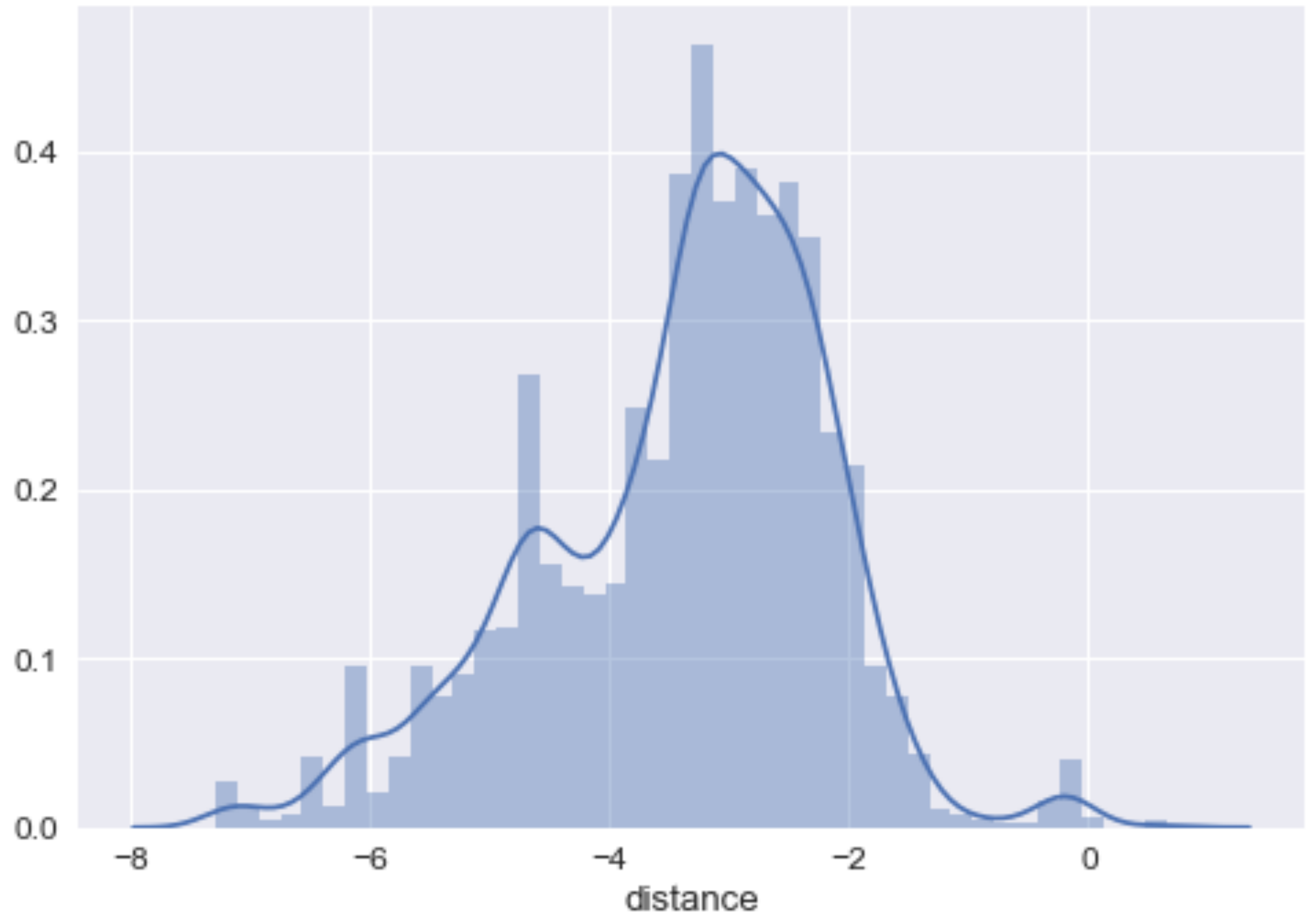




Exploratory Analysis – distance

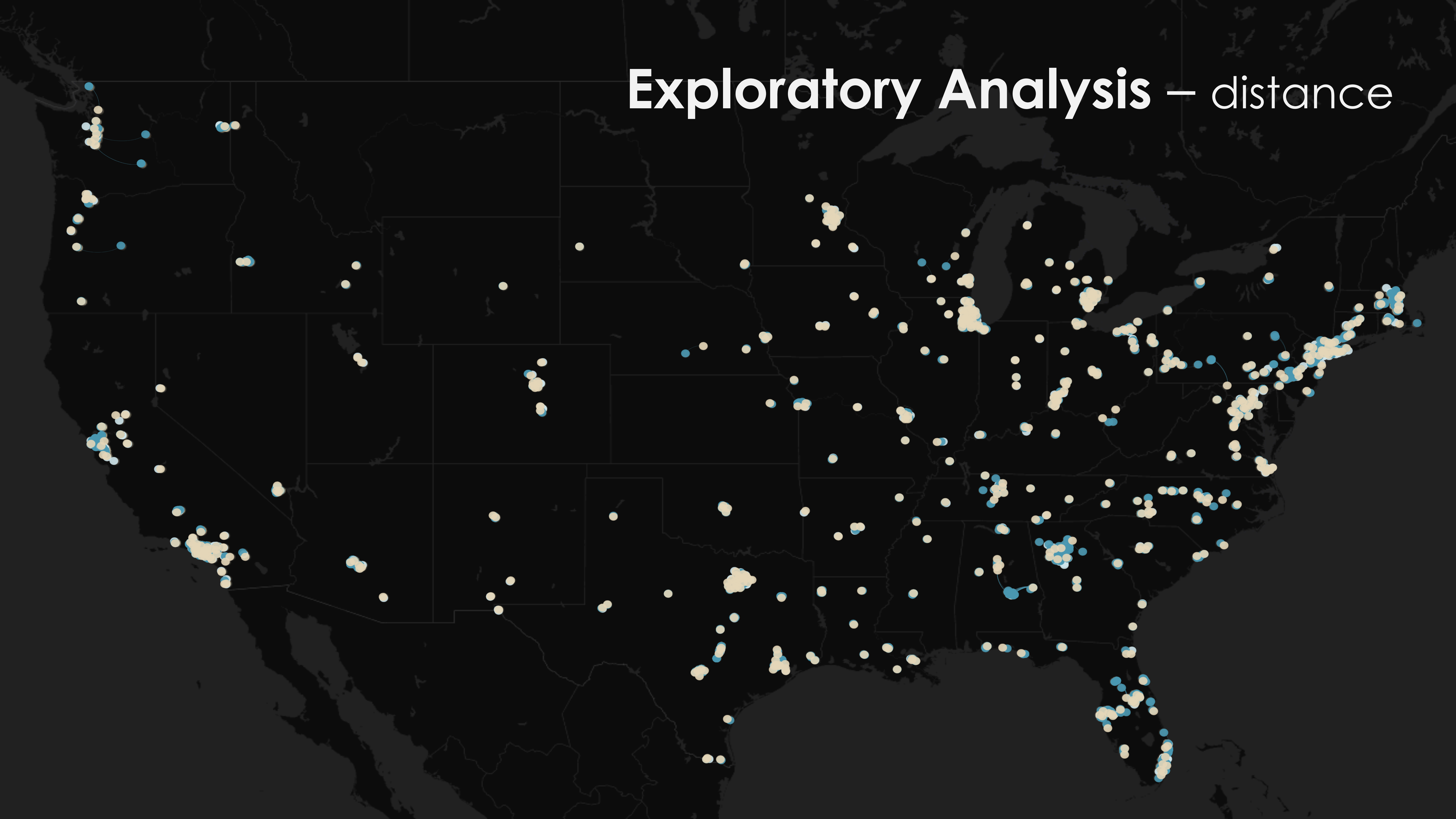


df_raw2['distance']

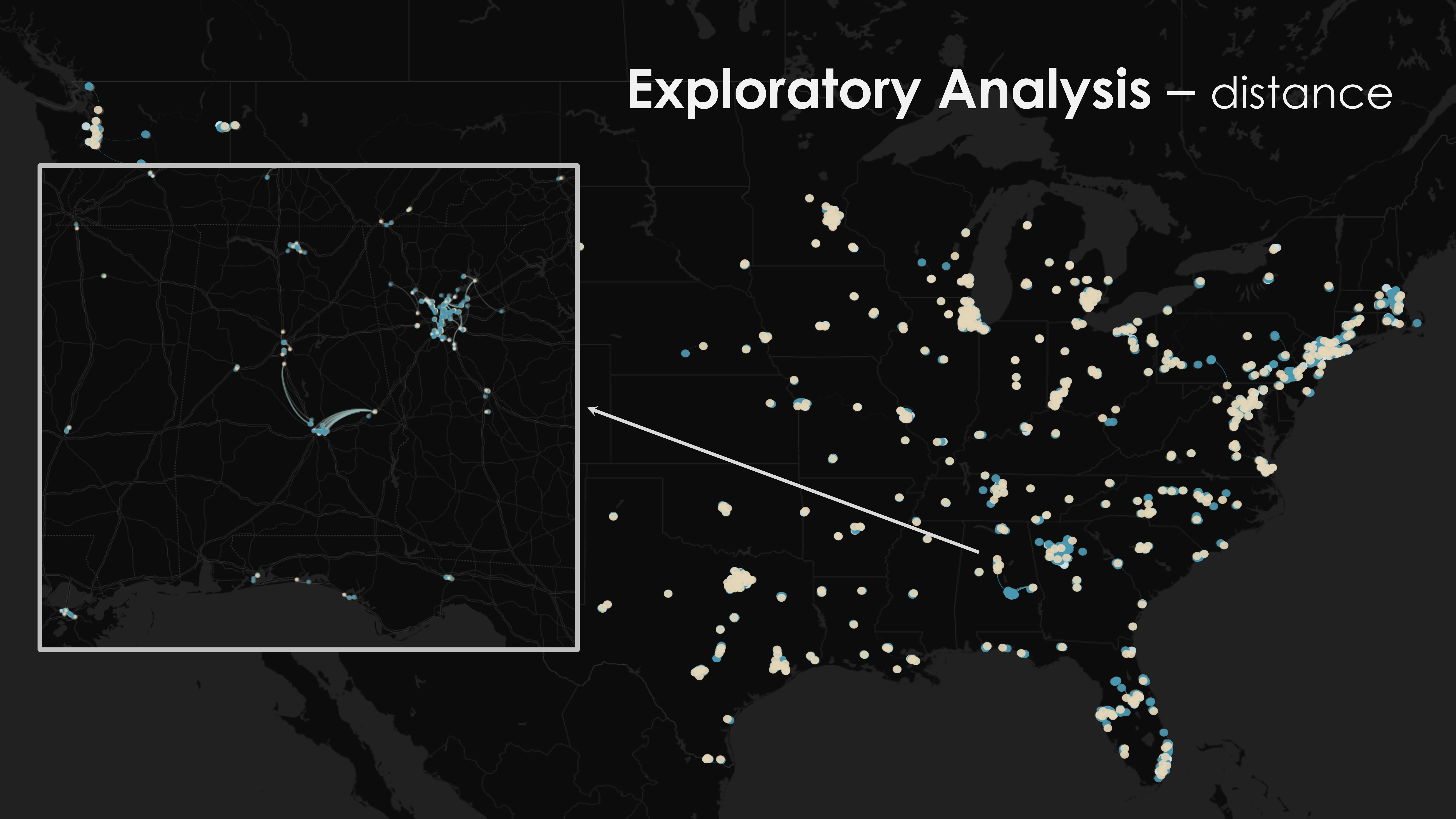


np.log(df_raw2['distance'])

Exploratory Analysis – distance



Exploratory Analysis – distance



Logistic Regression vs. Random Forest

Logistic Regression

- Train score*: 0.64605
- Test score: 0.59888

	precision	recall	f1-score	support
converted	0.97	**0.61**	0.75	1794
not-converted	0.06	**0.59**	0.10	70
avg / total	0.94	**0.61**	0.85	1864

Random Forest

- Train score: 0.706094
- Test score: 0.608656
- Average AUC: 0.60188

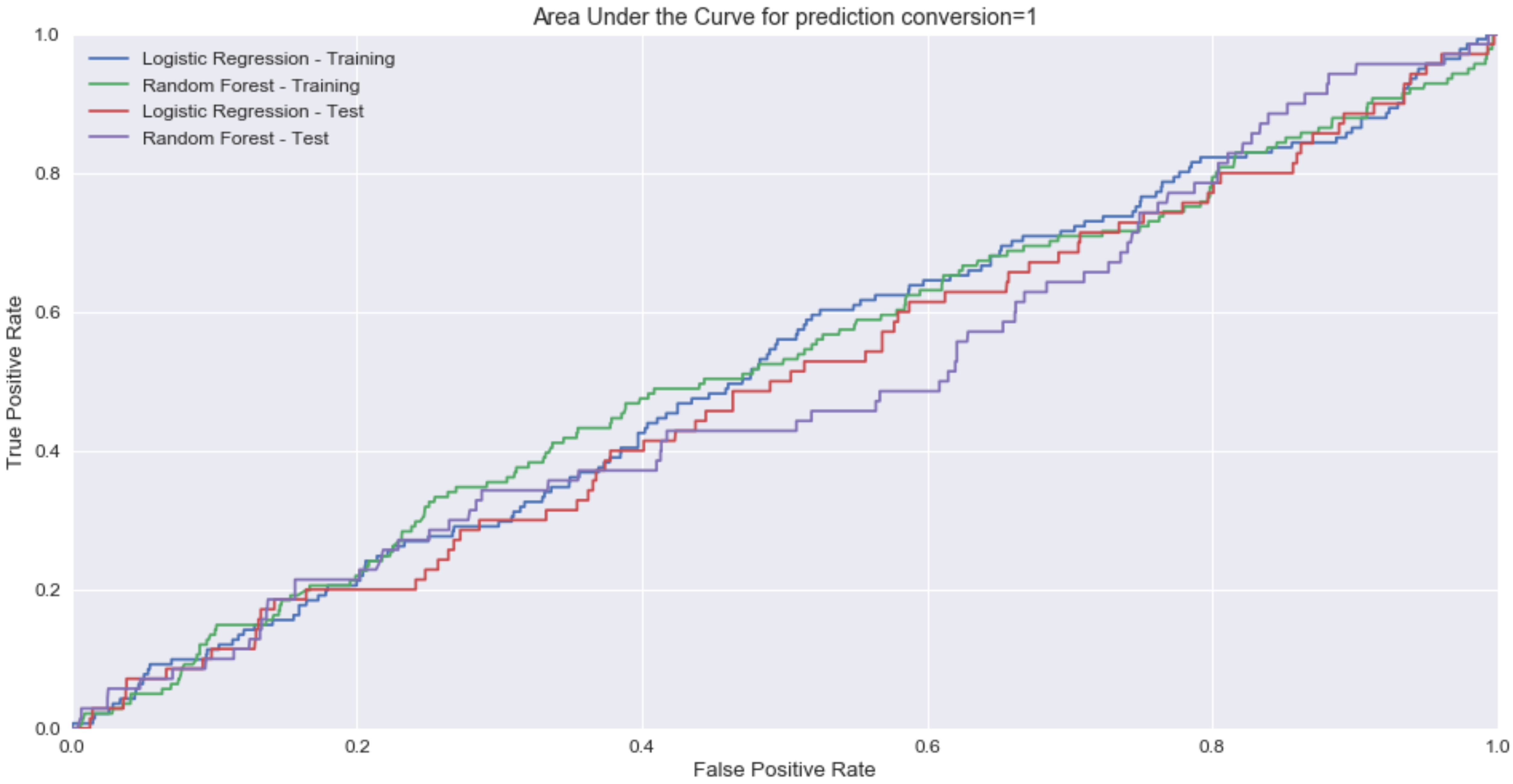
	precision	recall	f1-score	support
converted	0.97	**0.79**	0.87	1794
not-converted	0.07	**0.43**	0.13	70
avg / total	0.94	**0.78**	0.84	1864

**roc_auc_score rounded up to the nearest 5 decimal places.*

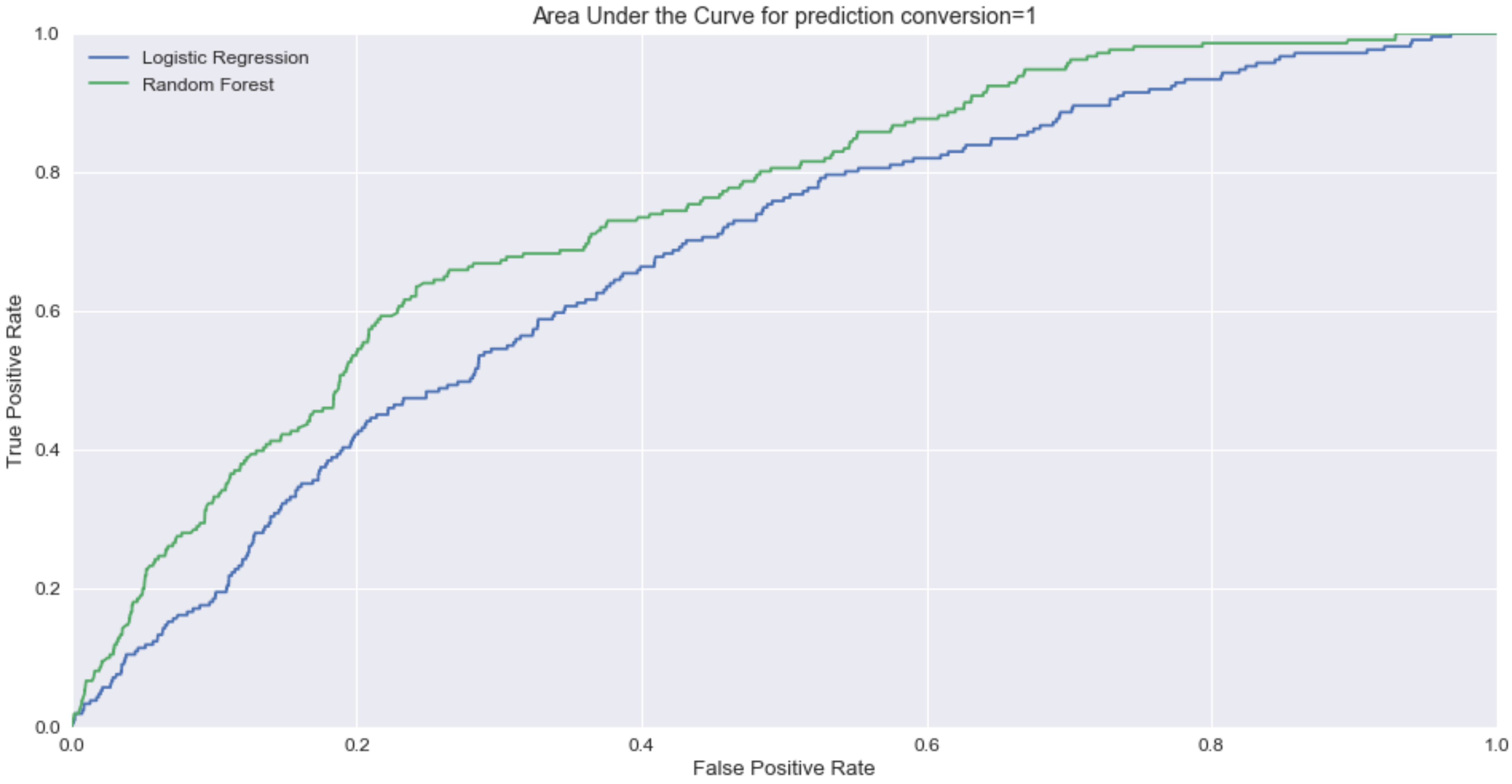
Model Evaluation Summary

	feature_set	penalty	scoring	train score	test score
gs1	1	L1	accuracy	0.50000	0.50000
gs2	1	L1	recall	0.64671	0.58654
gs3	1	L2	recall	0.64564	0.57504
gs4	2	L1	recall	0.64635	0.59146
gs5	3	L1	recall	0.64605	0.59888
gs6	3	L1	roc_auc	0.50000	0.50000
gs7	3	L1	precision	0.64539	0.59090
gs8	4	L1	recall	0.64713	0.58654
gs9	4	L1	precision	0.64713	0.58654
gs10	5	L1	recall	0.64713	0.58654
gs11	5	L2	recall	0.65218	0.57114
gs12	6	L1	recall	0.63097	0.56881
gs13	6	L1	precision	0.63138	0.56222
gs14	6	L2	recall	0.63017	0.56417
gs15	7	L1	recall	0.62976	0.56334
gs16	8	L1	recall	0.63127	0.56473
clf_rf2	3	NA	recall	0.74897	0.58761
gs17	10	L1	recall	0.646851	0.587100
gs18	11	L1	recall	0.650695	0.597205
gs19	12	L1	recall	0.645776	0.598599
clf_rf4	3	NA	recall	0.706094	0.608656

Model Evaluation – cont.



Model Evaluation – cont.



Feature Evaluation

Logistic Regression

Features	Coefficients
relationship_Married	-0.461641
job_Retired	-0.292666
dow_6	-0.217327
education_Finished High School	0.217294
relationship_Separated/Divorced	-0.209348
education_Finished College	0.19846
dow_0	-0.173468
dow_1	0.164452
education_Post Graduate Study	0.161412
office	0.160516

Random Forest

Features	Importance Score
age	0.114709
distance	0.114333
relationship_Married	0.069006
office	0.065727
dow_6	0.05288
billboard	0.049157
education_Some College	0.039842
dow_0	0.03294
education_Finished College	0.032064
relationship_In a relationship	0.0257

Next steps

- Explore other features, including conjunction features
- Refine model using data from other campaigns
- Add feedback features and apply Bayesian logistic regression to all campaign data
- Work with engineering to deploy the model to production