DATA SCIENCE & STEM SALARIES

OUR DATASET

Our dataset includes 62,000 salary records from top companies, scraped from levels.fyi

Nufterical VariableS:

The features we used included numerical variables such as timestamps, years of experience, and years worked at the company.

CaTegorical VariableS:

Categorical variables included company, level, title, location, education, race, and gender

PROJECT GOALS

QuEStIOnS tO Be AnSwereD:

1. What are the highest paying jobs in the tech industry?
2. Which companies pay the highest salaries?
3. Do men earn more than women in the tech industry?
4. Where are the tech jobs located?
5. What factors are important in determining income?

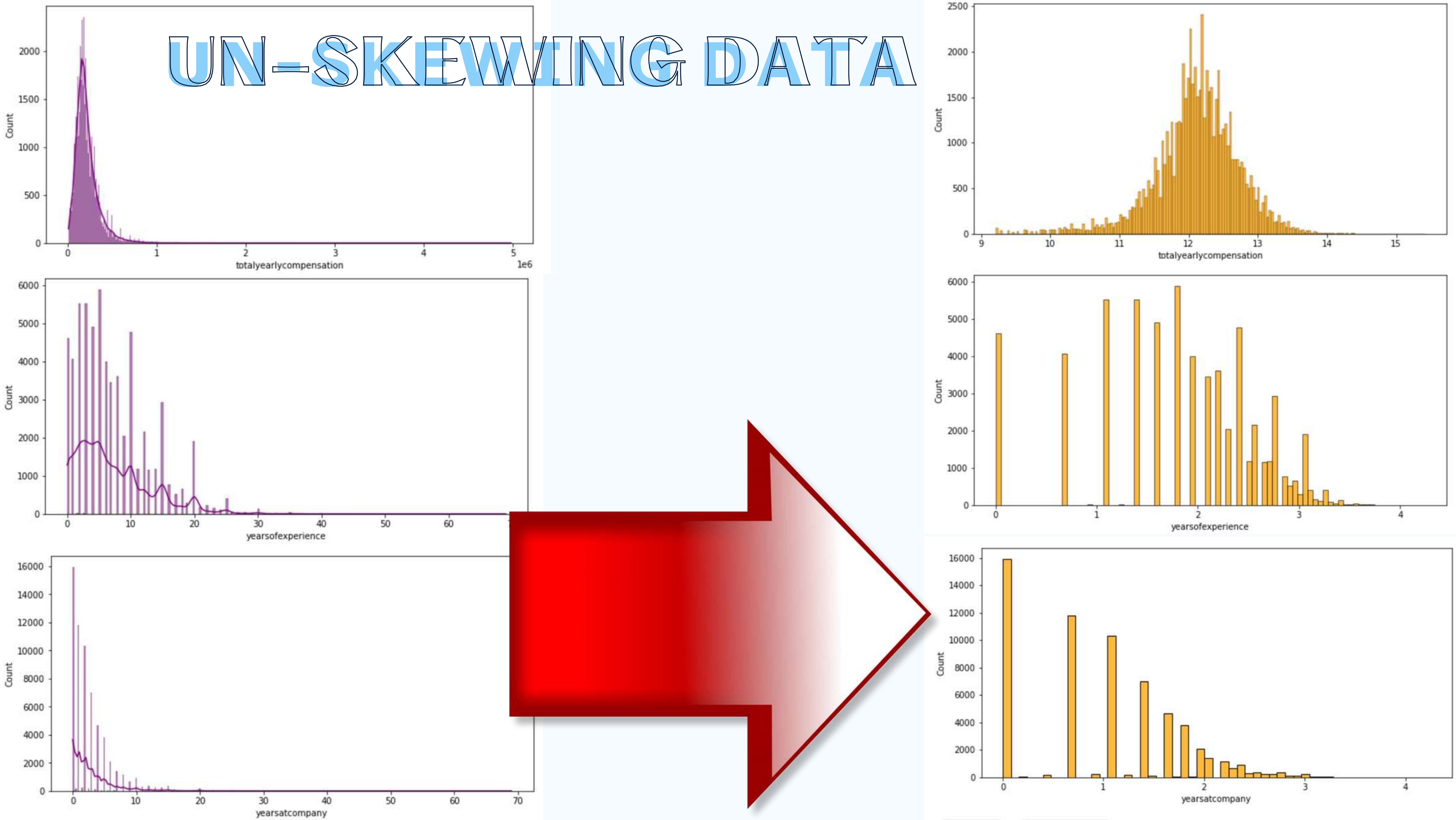
ImPOrtAnCe Of tHe PrOBleM:

Jobs and incomes affect everyone's lifestyles.

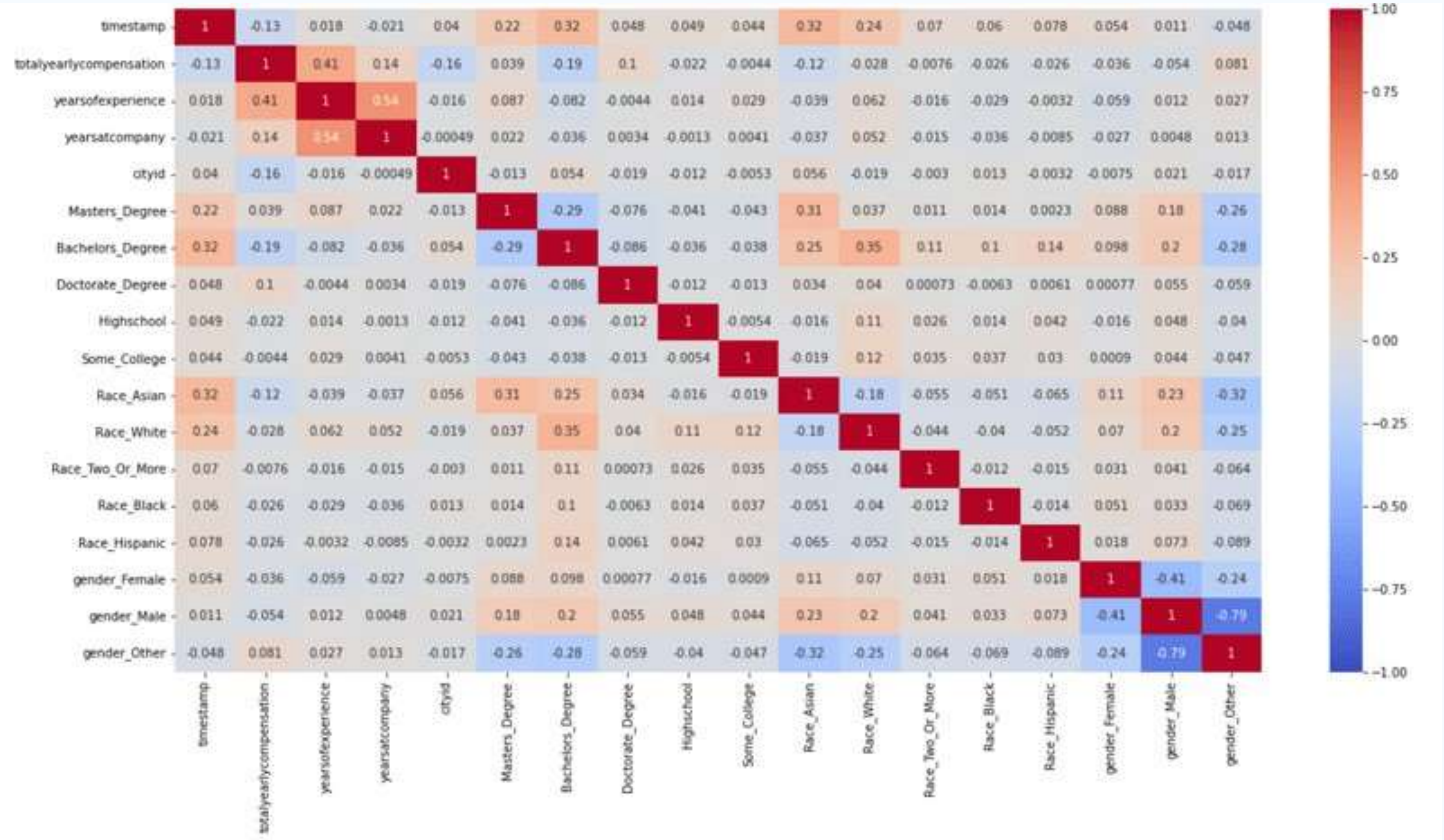
WHY we PICkeD tHIS:

Many of us will be joining the workforce soon as well

UN-SKEWING DATA



PATTERNS



Patterns found in the dataset

As expected, high degrees such as master's or doctorates correlate with high pay, as well as years of experience and at the company.

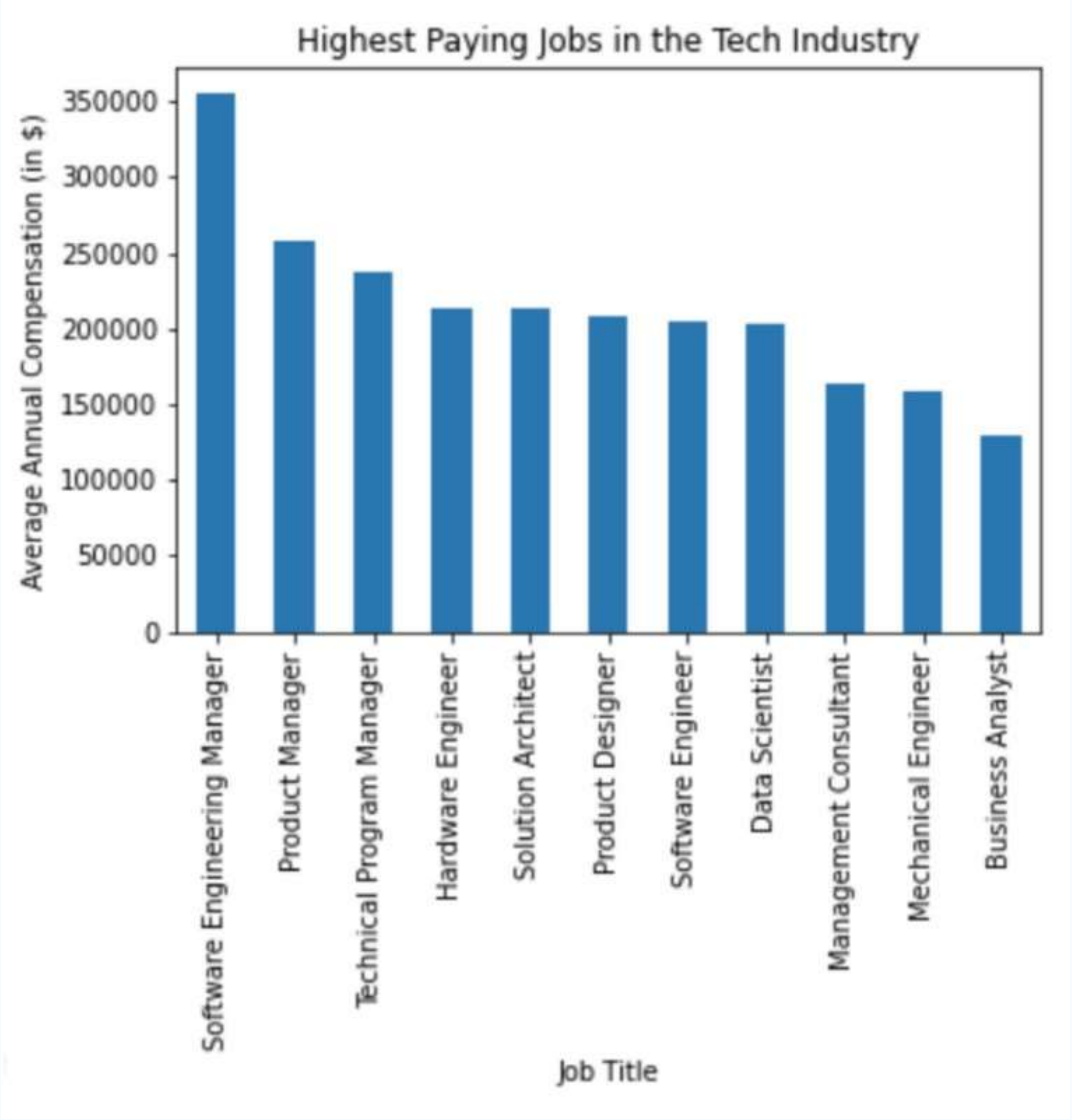
Any abnormalities found

Strangely, bachelor's degrees had a negative correlation, as well as every single race.

How they related to final goals of the project

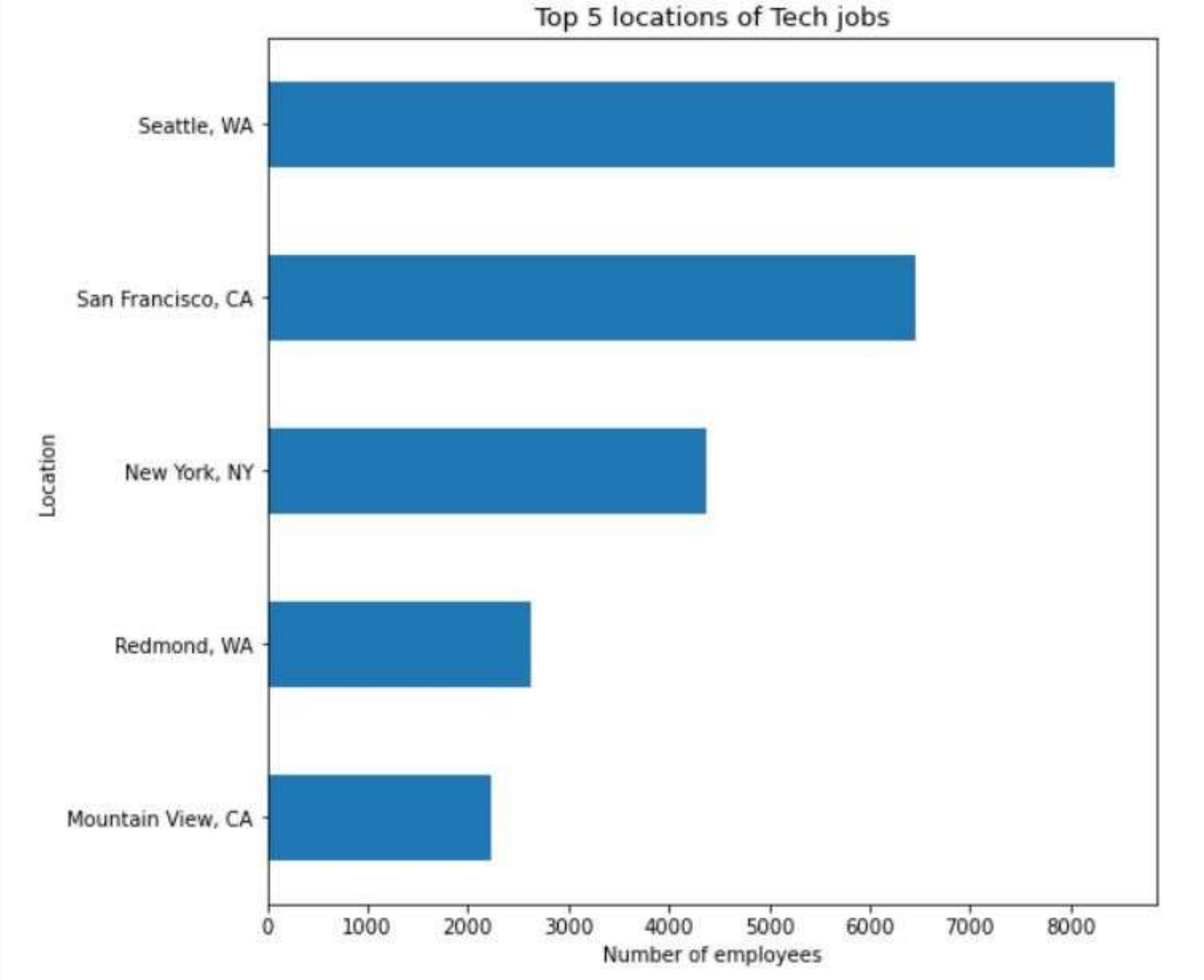
We want to gain a basic understanding of the data beforehand, and also try to reason out why some data is not behaving as expected.

HIGHEST PAYING JOBS



The highest (5) paying jobs in the tech industry include Software Engineer Manager, Product Manager, Technical Program Manager, Hardware Engineer, and Solution Architect.

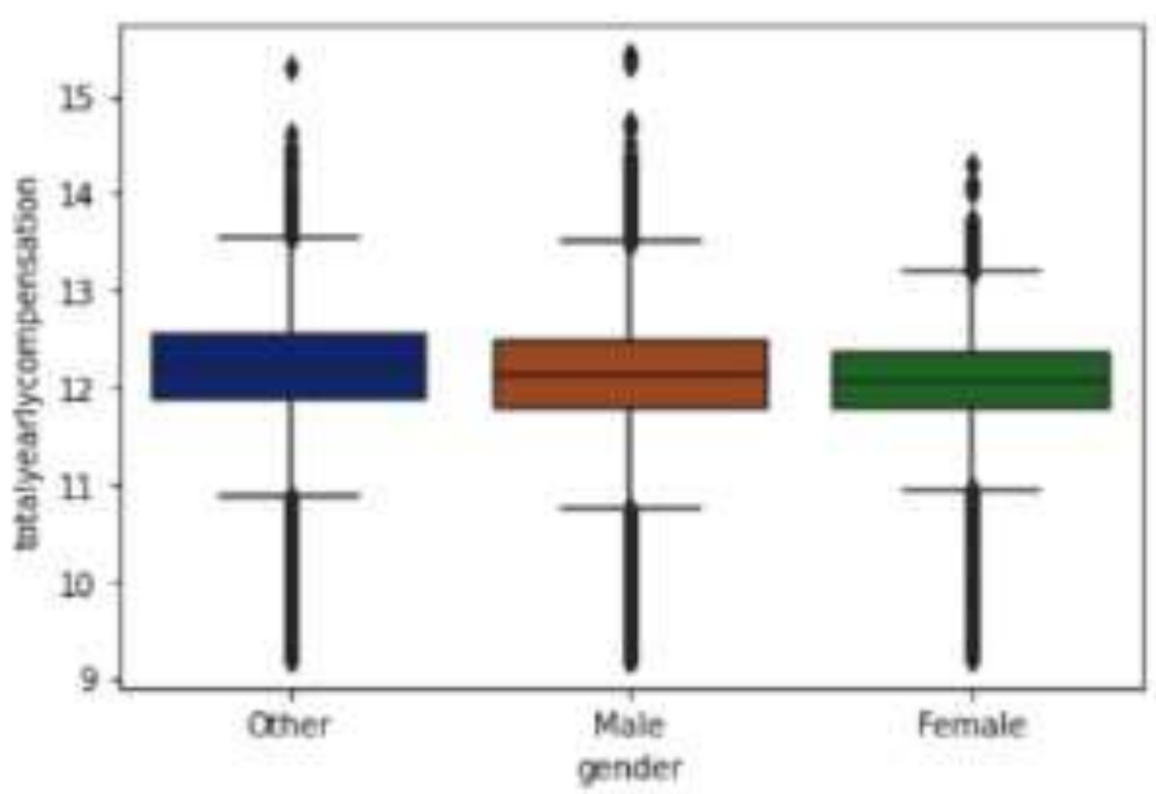
WHERE ARE TECH JOBS LOCATED?



The bar plot shows the top 5 locations for tech jobs. Seattle, WA has the highest number of tech jobs with over 8000 employees in the tech field.

DO MEN EARN MORE THAN WOMEN?

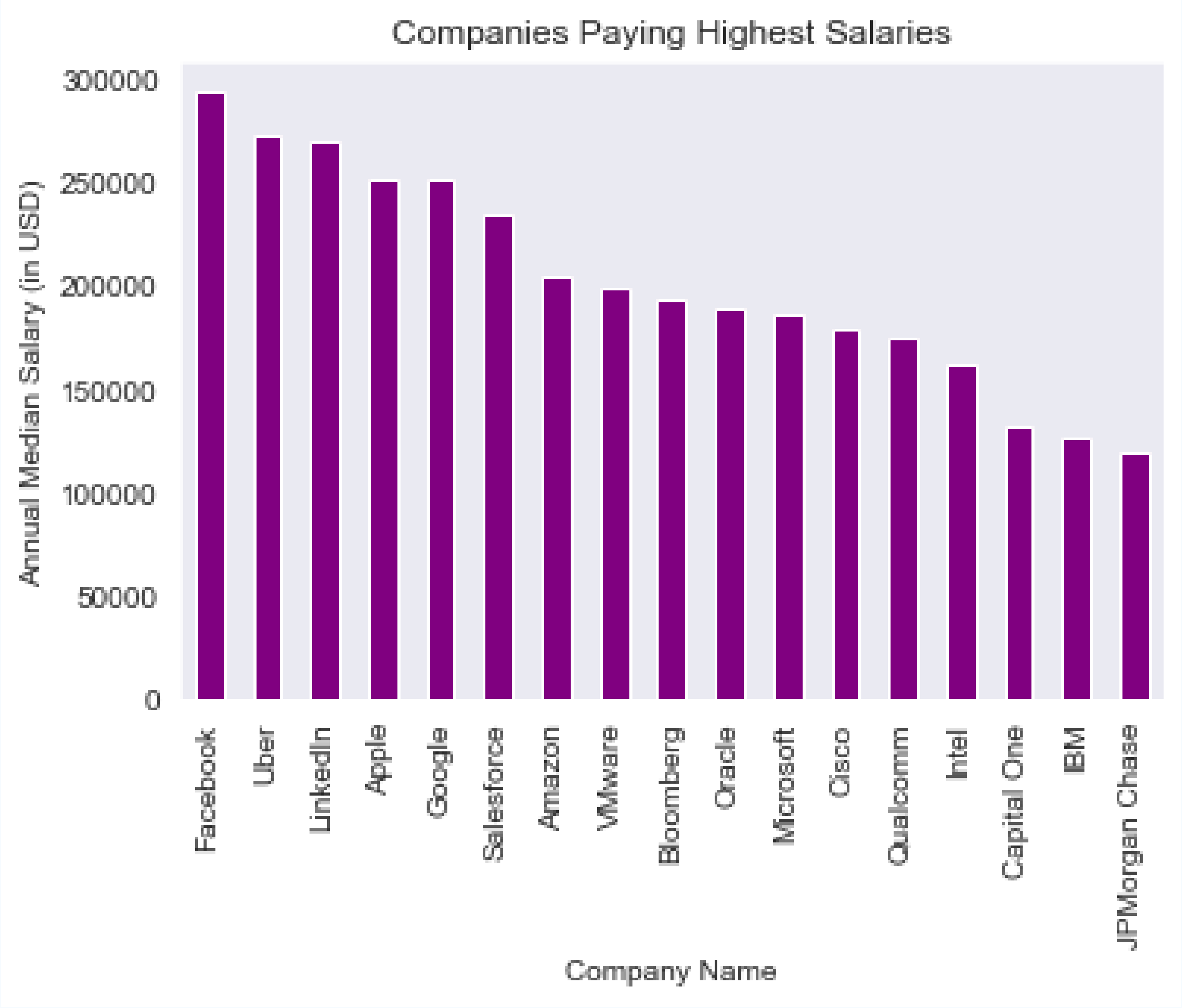
Figure 3



gender	
Female	12.051433
Male	12.084472
Other	12.186310

Yes, by about 3.3% on average

WHAT COMPANIES PAY THE HIGHEST SALARIES



Based on our analysis of companies with more than 500 employees, the highest annual median salaries in the tech industry range from $130k - $290k. Facebook offers the highest annual median salary followed by Uber, LinkedIn, Apple, Google and Salesforce in the same order.

SOLUTIONS & INSIGHTS

LINEAR REGRESSION

LINEAR REGRESSION

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.522e+05	3815.923	39.891	0.000	1.45e+05	1.6e+05
company[T.Apple]	7.66e+04	2913.733	26.288	0.000	7.09e+04	8.23e+04
company[T.Facebook]	1.337e+05	2484.709	53.812	0.000	1.29e+05	1.39e+05
company[T.Google]	5.221e+04	2375.685	21.975	0.000	4.75e+04	5.69e+04
company[T.Microsoft]	-2.266e+04	2506.566	-9.040	0.000	-2.76e+04	-1.77e+04
company[T.other]	-1.57e+04	1545.994	-10.153	0.000	-1.87e+04	-1.27e+04
title[T.Hardware Engineer]	6798.2977	3259.907	2.085	0.037	408.870	1.32e+04
title[T.Product Designer]	-2.194e+04	3613.801	-6.070	0.000	-2.9e+04	-1.49e+04
title[T.Product Manager]	2.469e+04	2764.880	8.931	0.000	1.93e+04	3.01e+04
title[T.Software Engineer]	9644.0908	2278.712	4.232	0.000	5177.808	1.41e+04
title[T.Software Engineering Manager]	9.604e+04	2949.700	32.560	0.000	9.03e+04	1.02e+05
title[T.other]	-2.352e+04	2741.122	-8.580	0.000	-2.89e+04	-1.81e+04
location[T.New York, NY]	-3.437e+04	3052.066	-11.263	0.000	-4.04e+04	-2.84e+04
location[T.Redmond, WA]	-3.086e+04	3984.150	-7.746	0.000	-3.87e+04	-2.31e+04
location[T.San Francisco, CA]	1.907e+04	2914.870	6.541	0.000	1.34e+04	2.48e+04
location[T.Seattle, WA]	-2.924e+04	2944.720	-9.931	0.000	-3.5e+04	-2.35e+04
location[T.other]	-8.503e+04	2650.352	-32.081	0.000	-9.02e+04	-7.98e+04
yearsofexperience	6.811e+04	677.247	100.562	0.000	6.68e+04	6.94e+04
yearsatcompany	-9209.9836	706.941	-13.028	0.000	-1.06e+04	-7824.377
Masters_Degree	-5318.3317	1105.472	-4.811	0.000	-7485.060	-3151.603
Bachelors_Degree	-2.244e+04	1202.123	-18.667	0.000	-2.48e+04	-2.01e+04
Doctorate_Degree	6.176e+04	2727.607	22.644	0.000	5.64e+04	6.71e+04
Highschool	-3.503e+04	6406.670	-5.468	0.000	-4.76e+04	-2.25e+04
Some_College	-2.476e+04	6077.139	-4.075	0.000	-3.67e+04	-1.29e+04

```
result.params
```

```
Intercept                                 152221.155266
company[T.Apple]                           76596.130365
company[T.Facebook]                       133707.869235
company[T.Google]                          52206.161374
company[T.Microsoft]                      -22658.905666
company[T.other]                          -15697.168215
title[T.Hardware Engineer]                  6798.297744
title[T.Product Designer]                 -21936.315175
title[T.Product Manager]                   24694.343459
title[T.Software Engineer]                  9644.090771
title[T.Software Engineering Manager]      96041.808834
title[T.other]                            -23519.531263
location[T.New York, NY]                  -34374.677383
location[T.Redmond, WA]                   -30861.423360
location[T.San Francisco, CA]              19066.588945
location[T.Seattle, WA]                   -29242.678675
location[T.other]                         -85027.208606
yearsofexperience                          68105.028954
yearsatcompany                             -9209.983639
Masters_Degree                             -5318.331653
Bachelors_Degree                          -22440.016113
Doctorate_Degree                           61764.016158
Highschool                                -35029.135584
Some_College                              -24764.771930
```

```
result.rsquared
```

```
0.36594163129201596
```

SOLUTIONS
& INSIGHTS
RIDGE AND LASSO

RIDGE REGRESSION

```
Ridge Regression:

Training Data
RMSE:  0.4696092898831499
R-Squared:  0.4118139387560622

Testing Data
RMSE:  0.4728856171745504
R-Squared:  0.4179828462694547
```

```
Best Hyperparameters: {'alpha': 0.0015264179671752333}
```

	Columns	Coef
11	yearsofexperience	0.258314
14	company_Facebook	0.098329
15	location_San Francisco, CA	0.089152
10	company_Apple	0.064012
24	title_Software Engineering Manager	0.056461
0	Doctorate_Degree	0.051863
7	company_Google	0.045562
23	location_Mountain View, CA	0.038543
30	location_Seattle, WA	0.034031
3	location_New York, NY	0.025509

LASSO REGRESSION

```
Lasso Regression:

Training Data
RMSE:  0.4696377190638314
R-Squared:  0.4117427214513079 7
```

```
Lasso Regression:

Training Data
RMSE:  0.4696377190638314
R-Squared:  0.41174272145130797

Testing Data
RMSE:  0.4729868844523289 5
R-Squared:  0.41773354452929046
```

```
Best Hyperparameters: {'alpha': 0.001}
```

```
Coefficients:
 [ 0. -0. -0.  0.  0. -0.  0.  0.  0.  0. -0. -0.  0.  0. -0.  0.  0. -0.
 -0. -0.  0. -0. -0. -0.  0.  0. -0.  0.  0.  0.  0. -0. -0.  0. -0.]
```

```
Lasso Regression (lambda = 0.1):

Training Data
RMSE:  0.531292723164083
R-Squared:  0.24714887358968918

Testing Data
RMSE:  0.5370713385018615
R-Squared:  0.24926350727637026
```

SOLUTIONS & INSIGHTS NEAREST NEIGHBORS

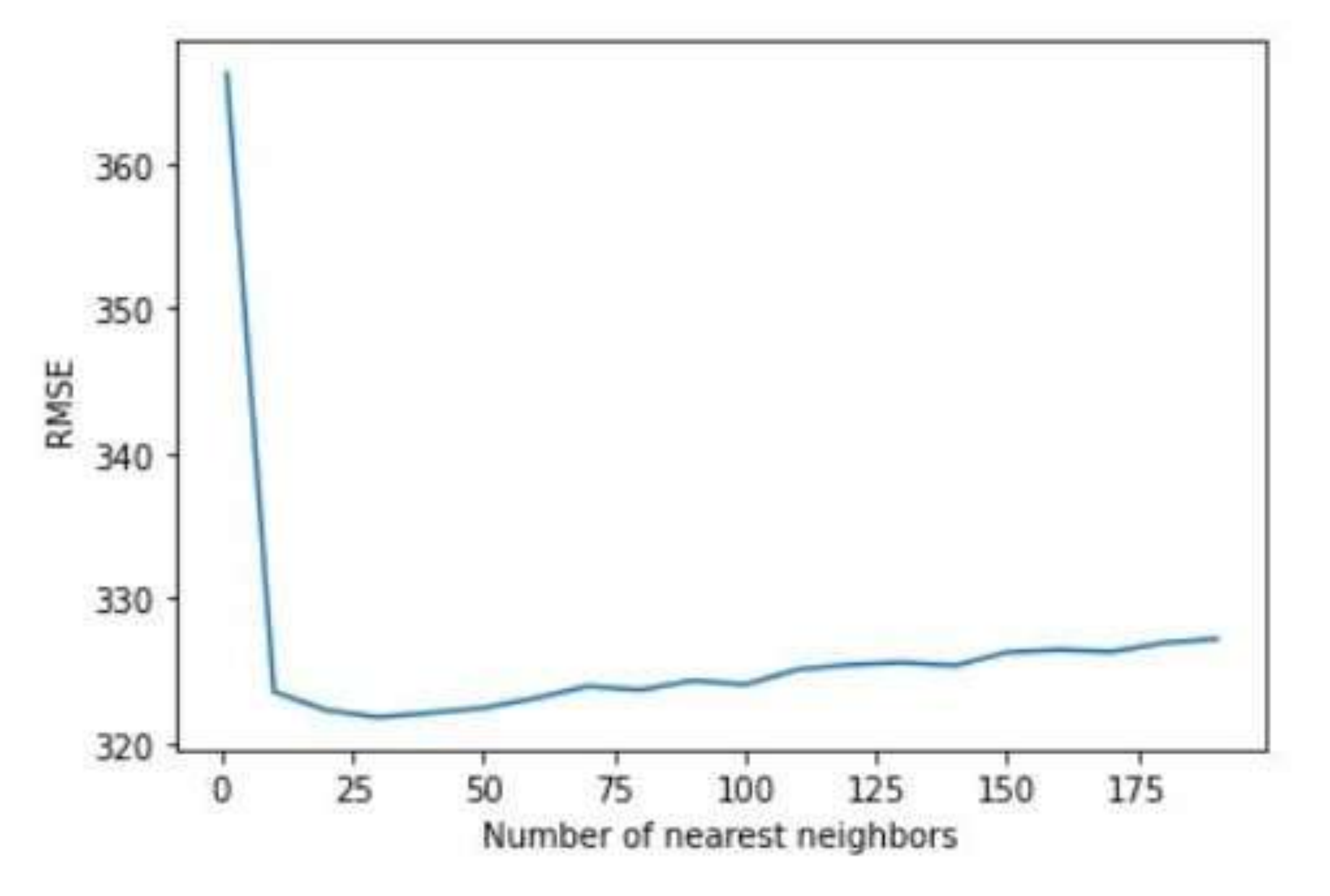
Number of neighbors vs RMSE

Output Variable: Total yearly compensation in USD

Input Predictors: All the variables were used as predictors

RMSE obtained: 321.78

Best K value: 30



SOLUTIONS & INSIGHTS RANDOM FOREST

FEATURE ENGINEERING

	totalyearlycompensation	yearsofexperience	yearsatcompany
0	127000	1.5	1.5
1	100000	5.0	3.0
2	310000	8.0	0.0
3	372000	7.0	5.0
4	157000	5.0	3.0

5 rows × 84112 columns

	timestamp	yearsofexperience	yearsatcompany	cityid
0	2017	0.916291	0.916291	7392
1	2017	1.791759	1.386294	7419
2	2017	2.197225	0.000000	11527
3	2017	2.079442	1.791759	7472
4	2017	1.791759	1.386294	7322

5 rows × 212 columns

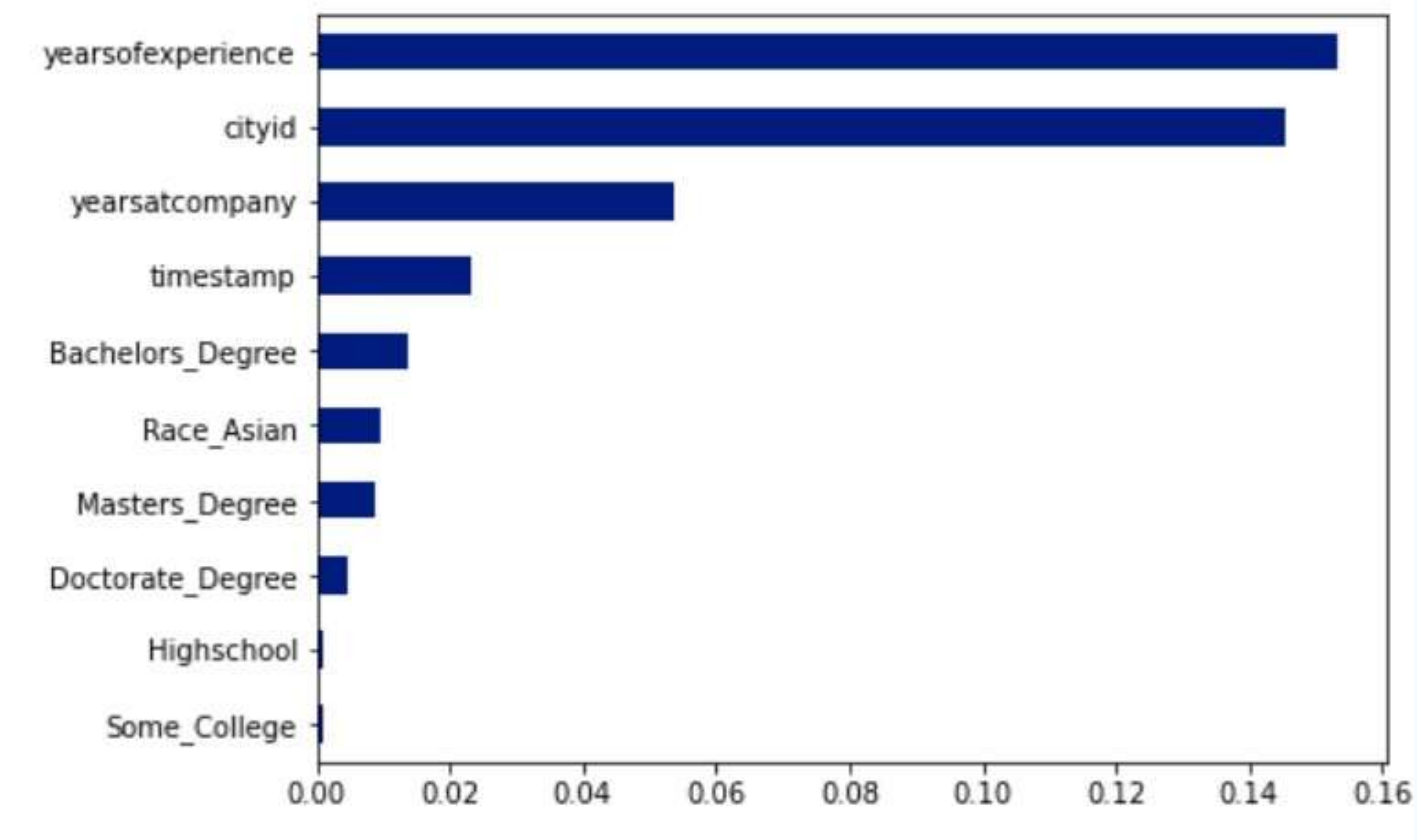
Random Forest is extremely slow for a large number of features.
A naive get_dummies() returned 84k features!

Instead, the year was extracted from timestamps,
the state or country from location instead of individual cities,
and for other categorical features such as company or tag,
anything with less than 500 features (<1% of 62 k total data points) were lumped into an "other category.

As shown, this compresses the data by about 40 times.

RESULTS

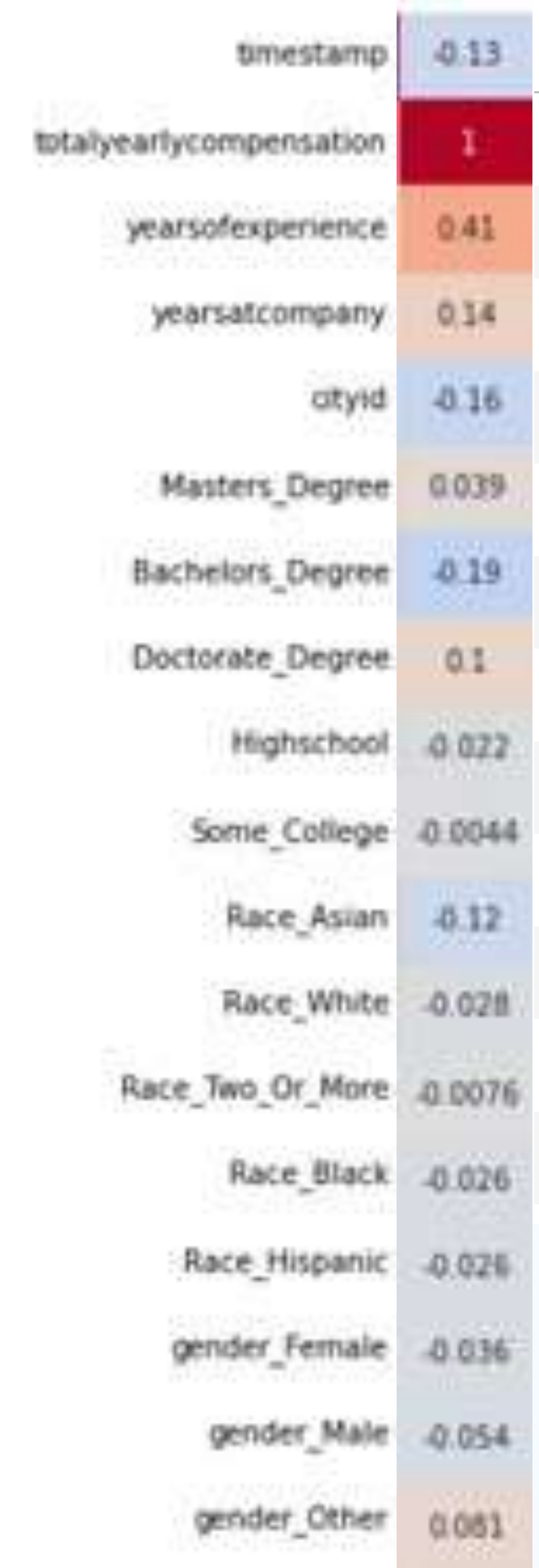
	MSE	R2
Values	0.097692	0.737735



After tuning the number of estimators, maximum features considered per split, and the minimum number of samples allowed per leaf node, a RMSE of 0.312 was achieved.

It seems that the two most important features by far are the years of experience someone has, as well as the city they work in.
The number of years worked at the company also has a bit of an effect

RESULTS



	Columns	Coefs
0	yearsofexperience	0.344985
1	yearsatcompany	-0.080529
2	timestamp	-0.001346
3	Bachelors_Degree	-0.047461
4	Race_Asian	-0.045527
5	Masters_Degree	0.009562
6	Doctorate_Degree	0.271742
7	Highschool	-0.039831
8	Some_College	-0.085511
9	cityid_0	-0.107074
10	cityid_11	0.101949
11	cityid_12	0.157046

Random Forest does show feature importance, but lacks explainability of the relationship for each.

Looking back at the correlation matrix, as well as fitting a simple linear regression on the most important features, we have a better understanding of the direction and magnitude of how each column affects pay.

SOLUTIONS & INSIGHTS
GRADIENT BOOSTING

FEATURE ENGINEERING

- All predictors with string data types are converted to uppercase and transformed using label encoder.
- Initial **Histogram Gradient Boosting Regressor** is built with hyper parameters :

1. number of boosting stages : **500**
2. learning rate : **0.01**

```
# Hyperparameters for GradientBoostingRegressor

gbr_params = {  'max_iter' : 500,
                'learning_rate': 0.01,
             }

# Create an instance of gradient boosting regressor
gbr = HistGradientBoostingRegressor(**gbr_params)

# Fit the model
gbr.fit(X_train, y_train)
```

```
HistGradientBoostingRegressor(learning_rate=0.01, max_iter=500)
```

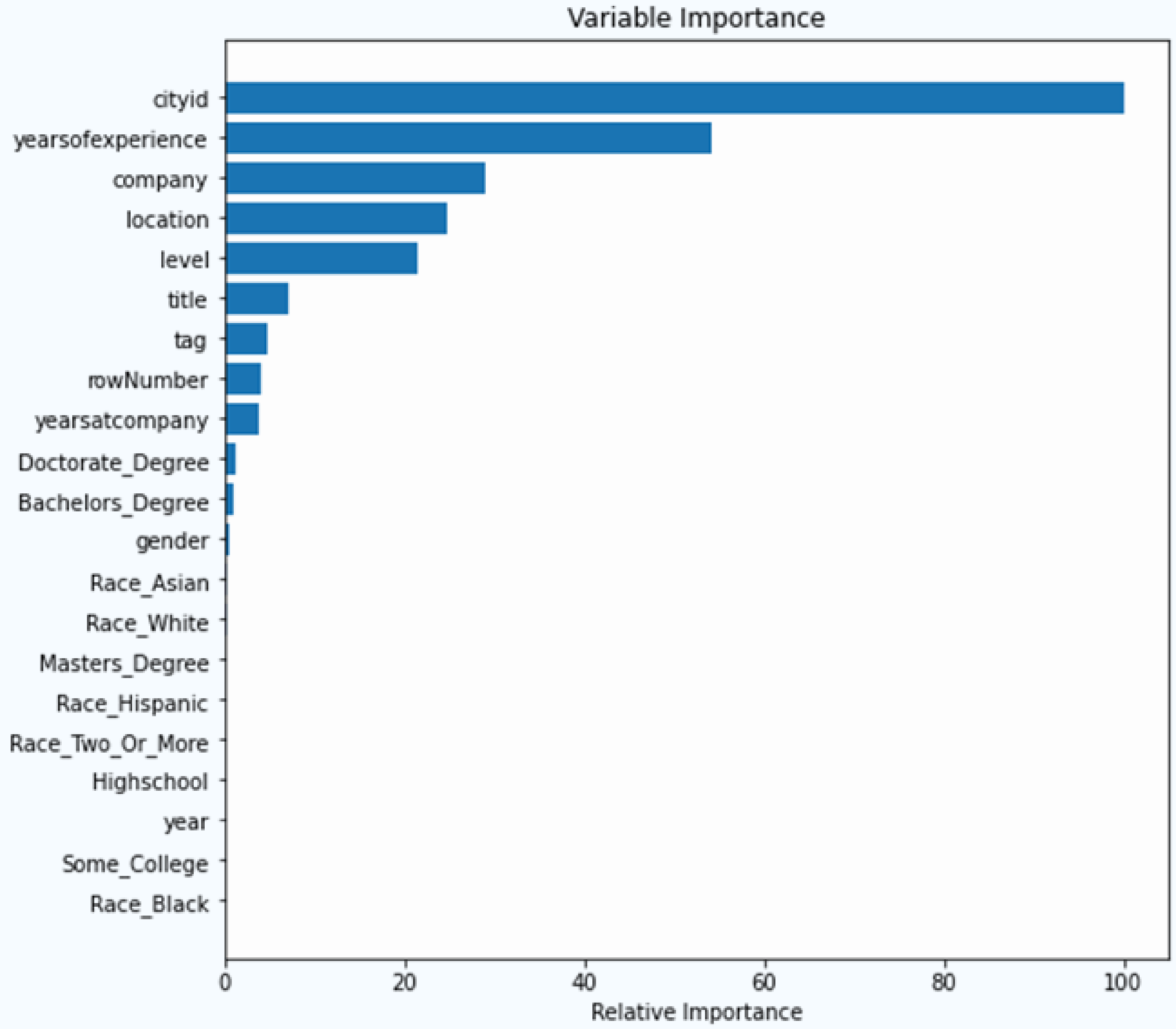
Result :

```
# Print Coefficient of determination R^2
print("R-squared: %.3f" % gbr.score(X_test, y_test))

# Create the mean squared error
mse = math.sqrt(mean_squared_error(y_test, gbr.predict(X_test)))
print("The root mean squared error (MSE) on test set: {:.4f}".format(mse))
```

```
R-squared: 0.713
The root mean squared error (MSE) on test set: 0.3277
```

FEATURE IMPORTANCE



Results of Relative Feature Importance

- **Relevant variables** : city id, years of experience, company, location, level, title, tag, years at company, Doctorate and Bachelor’s degree in the same order.
- **Gender** and **race** don’t play a relative significant role in determining the salary of the individual in the tech industry.
- However, a Master’s degree isn’t as important predictor for salary as compared to a Bachelor's or Doctorate degree !!

RESULTS

RandomizedSearchCV is used for hyperparameters tuning.

```
print(" Results from Randomized Search " )
print("\n Best estimator across ALL searched params:\n",search.best_estimator_)
print("\n Best score across ALL searched params:\n",search.best_score_)
print("\n Best parameters across ALL searched params:\n",search.best_params_)
```

```
 Results from Randomized Search

 The best estimator across ALL searched params:
 HistGradientBoostingRegressor(max_depth=7, max_iter=1000)

 The best score across ALL searched params:
 0.801445817690548

 The best parameters across ALL searched params:
 {'max_iter': 1000, 'max_depth': 7, 'learning_rate': 0.1}
```

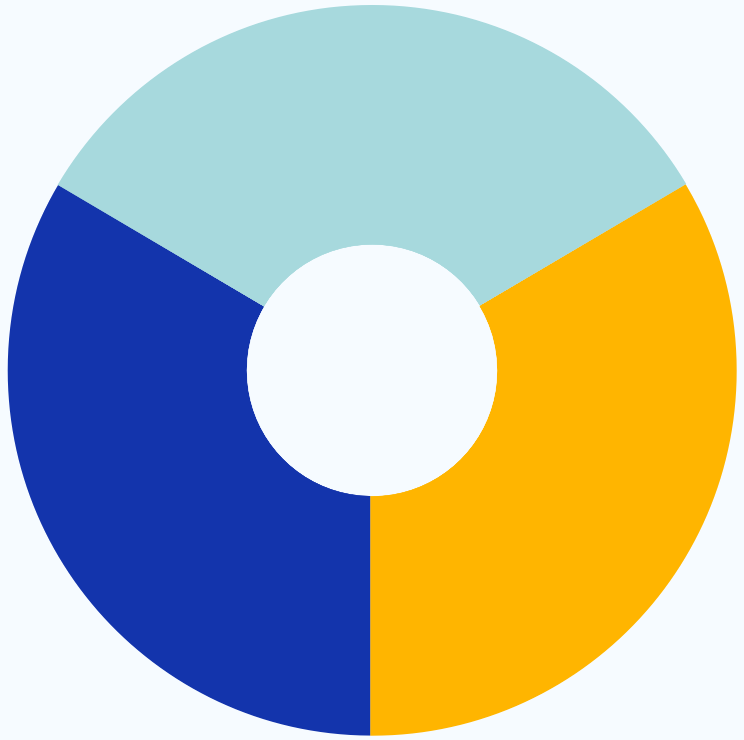
After Hyperparameter tuning :

```
# Print Coefficient of determination R^2
print("R-squared: %.3f" % gbr.score(X_test,y_test))

# Create the mean squared error
rmse = math.sqrt(mean_squared_error(y_test, gbr.predict(X_test)))
print("The mean squared error (RMSE) on test set: {:.4f}".format(rmse))
```

```
R-squared: 0.804
The mean squared error (RMSE) on test set: 0.2707
```

SOLUTIONS & INSIGHTS



Overall, the most important and consistent features when it came to determining pay were:

- Years of Experience
- Location

Other factors included:

- Years at company
- Company
- Level
- Title

SUMMARY OF OUR RESULTS

MODELS USED

Our models included linear regression, nearest neighbors, random forest, and boosting.

WHAT WE FOUND

The two most important features were years of experience and location.

WHAT SURPRISED US

We were surprised to see level of education absent from our list of important features

THANK YOU!