Seattle Airbnb Price Prediction

Karthik Garimella, Sandeep Alfred

MA 5790 – Fall 2024 – Team 12

Goal of the Study

- To Predict Prices of Airbnb Homestays in the city of Seattle.

Getting the data

- Downloaded the latest quarterly listings of houses from Airbnb.
- Added state and city names to the dataset whilst extracting the listing of each city.
- Filtered the cities to Seattle to get our dataset.

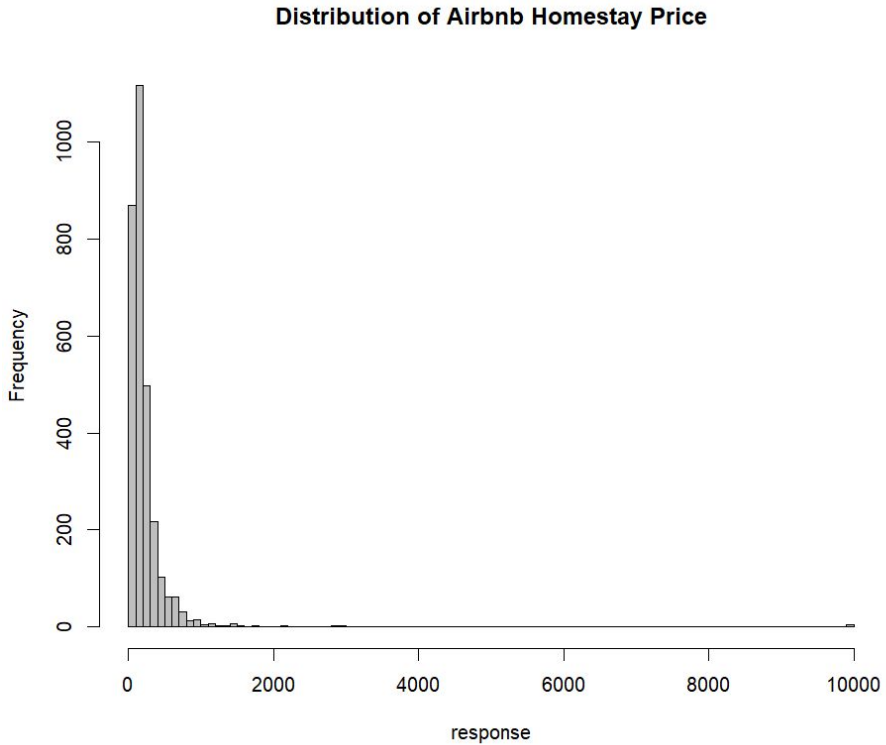
About The Data

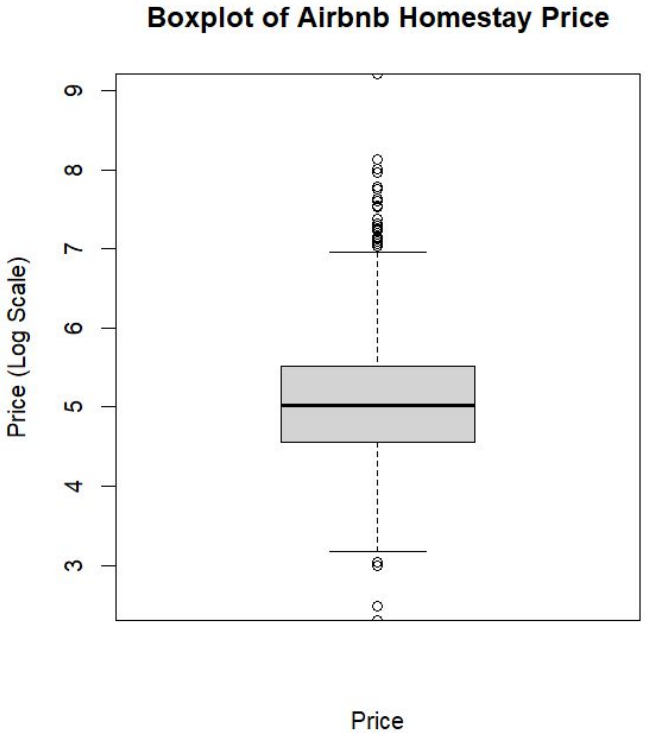
- Number of observations: **6442**
- Response Variable: **Price** (per night of stay)
- Predictor Count: **76**

Predictor Type	Count
Continuous	38
Categorical	27
Boolean	6
Date	5

Response variable Diagnostics

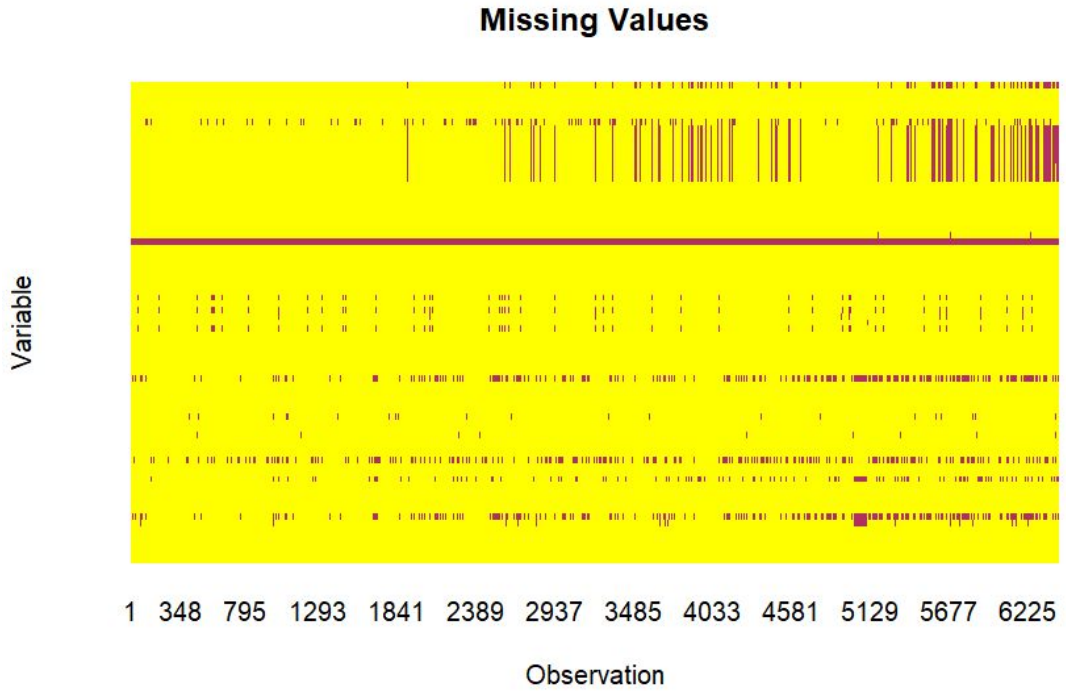
- The response variable is continuous and seems highly right skewed, there are also a decent number of outliers.





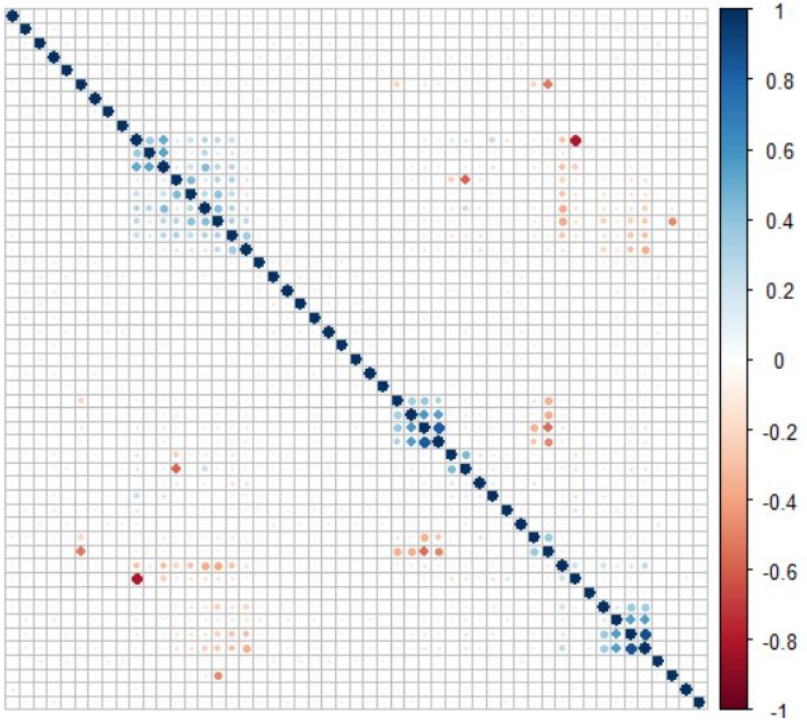
Preprocessing: Managing the Predictors

1. First **eliminated** the variables such as urls, description and the ones mostly containing null values and are repeated in different names, making the **predictors drop from 76 to 28.**
2. Removed rows with null values in the response variable.
3. Converted bedrooms, and all the month columns as categorical variables, and added the dummy variables, **increasing predictors from 28 to 71.**
4. KNN imputed all the remaining missing values.



Preprocessing: Near Zero Variance and Highly Correlated

- Eliminated the predictors with near zero variances, making the **predictors drop from 71 to 51.**
- Eliminated the predictors with high correlations (`>0.8`), making the **predictors drop from 51 to 48.**
- The dataset without the highly correlated predictors was only used with models which cannot handle highly correlated variables.



Preprocessing: Transformations

- Centered and scaled the data.
- Applied **BoxCox** transformation to improve the distributions of the predictors.
- Applied **spatial sign** transformation to reduce the impact of outliers.

	skewValues	Skewness
host_response_rate	-7.5006584	Highly Skewed
host_acceptance_rate	-3.6238753	Highly Skewed
host_is_superhost	-0.1421836	Approx. Symmetric
host_total_listings_count	4.1329698	Highly Skewed
host_identity_verified	-3.2906143	Highly Skewed
latitude	-0.2790336	Approx. Symmetric
longitude	-0.1545022	Approx. Symmetric
accommodates	1.6915679	Highly Skewed
bathrooms	5.0896143	Highly Skewed
beds	2.0089727	Highly Skewed
minimum_nights	7.1921129	Highly Skewed
maximum_nights	2.0686794	Highly Skewed
availability_30	1.1025193	Highly Skewed

	skewValues	Skewness
host_response_rate	-6.0007302	Highly Skewed
host_acceptance_rate	-2.8983224	Highly Skewed
host_is_superhost	-0.1568171	Approx. Symmetric
host_identity_verified	-3.3513378	Highly Skewed
latitude	-0.2348160	Approx. Symmetric
longitude	-0.1838311	Approx. Symmetric
bathrooms	1.7293419	Highly Skewed
beds	1.3969819	Highly Skewed
minimum_nights	2.1389492	Highly Skewed
maximum_nights	0.6992554	Moderately Skewed
availability_30	1.0435285	Highly Skewed
number_of_reviews	2.1930040	Highly Skewed
review_scores_value	-2.4502025	Highly Skewed

Principal Component Analysis

- PCA suggests 40 components to capture 95% of the variance.

```
Created from 6011 samples and 51 variables

Pre-processing:
  - centered (51)
  - ignored (0)
  - principal component signal extraction (51)
  - scaled (51)

PCA needed 40 components to capture 95 percent
of the variance
```

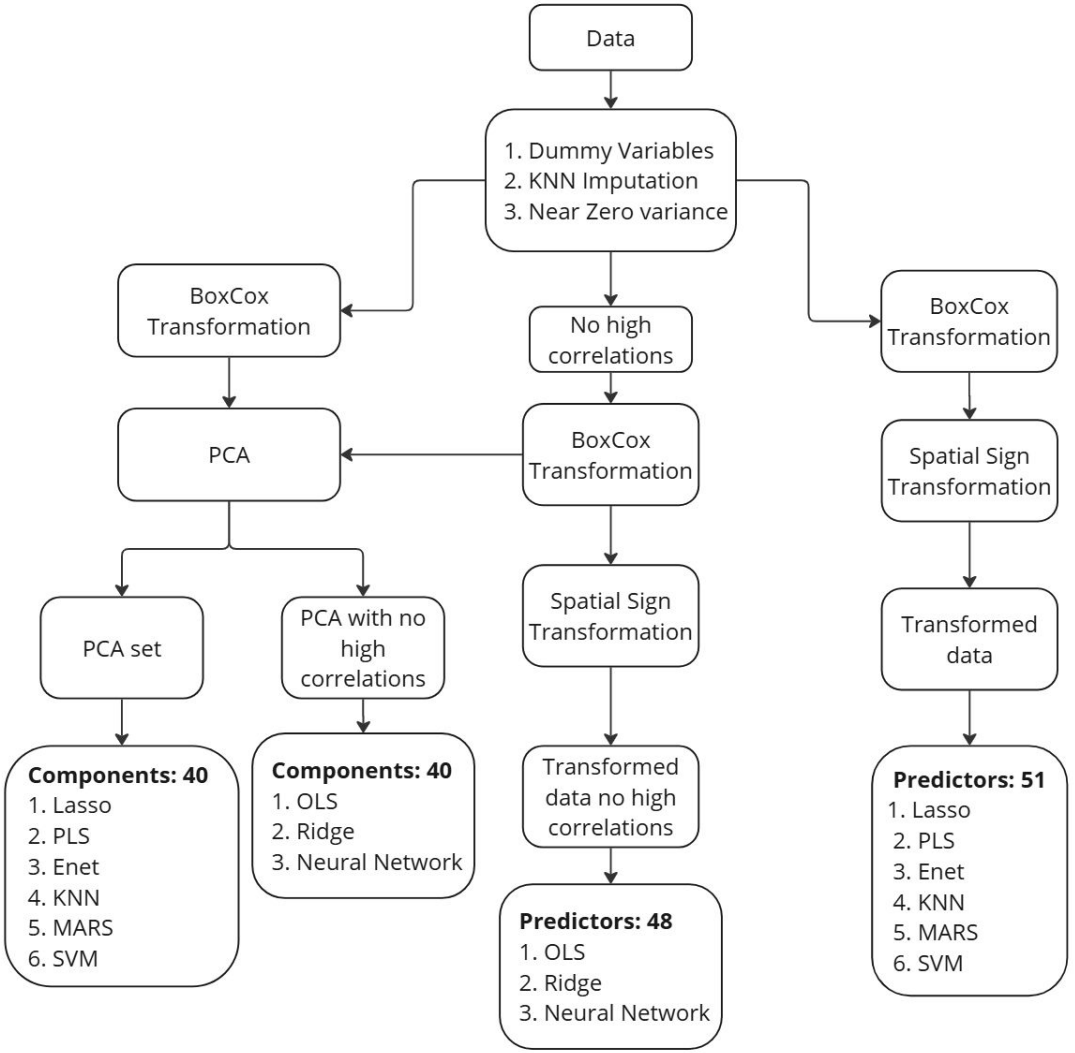
Dataset containing highly correlated variables

```
Created from 6011 samples and 48 variables

Pre-processing:
  - centered (48)
  - ignored (0)
  - principal component signal extraction (48)
  - scaled (48)

PCA needed 40 components to capture 95 percent
of the variance
```

Dataset not containing highly correlated variables

Modelling

- Split the data into 80% Train and 20% Test sets.
- Used 10 fold Cross-Validation resampling on all the models.
- Used Rsquared as the metric to choose the best parameters.



miro

Linear Models: Ordinary Linear Regression

Transformed data:

```
Linear Regression

4811 samples
  48 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4331, 4330, 4330, 4331, 4329, 4330, ...
Resampling results:

  RMSE      Rsquared   MAE
  169.4685  0.3553342  76.03205

Tuning parameter 'intercept' was held constant at a value
 of TRUE
```

PCA:

```
Linear Regression

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4331, 4331, 4331, 4329, 4330, ...
Resampling results:

  RMSE     Rsquared   MAE
  167.764  0.3795297  77.09713

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Linear Models: Ridge Regression

Transformed data:

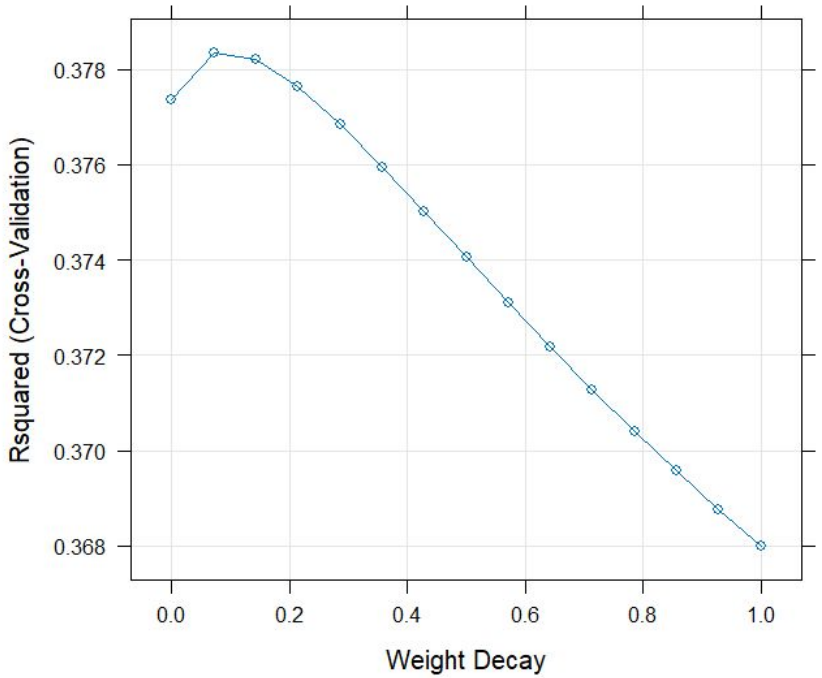
```
Ridge Regression

4811 samples
  48 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4331, 4331, 4331, 4329, 4330, ...
Resampling results across tuning parameters:

  lambda      RMSE      Rsquared   MAE
  0.00000000  167.9821  0.3773754  78.33443
  0.07142857  168.0211  0.3783392  78.56273
  0.14285714  168.2519  0.3782148  79.08590
  0.21428571  168.6023  0.3776428  79.78426
  0.28571429  169.0409  0.3768516  80.59061
  0.35714286  169.5481  0.3759523  81.49212
  0.42857143  170.1094  0.3750059  82.43888
  0.50000000  170.7135  0.3740476  83.43104
  0.57142857  171.3513  0.3730982  84.45061
  0.64285714  172.0154  0.3721701  85.47817
  0.71428571  172.6996  0.3712703  86.52379
  0.78571429  173.3988  0.3704025  87.56908
  0.85714286  174.1086  0.3695685  88.61320
  0.92857143  174.8255  0.3687688  89.65993
  1.00000000  175.5465  0.3680029  90.69664

Rsquared was used to select the optimal model using
 the largest value.
The final value used for the model was lambda = 0.07142857.
```

PCA:

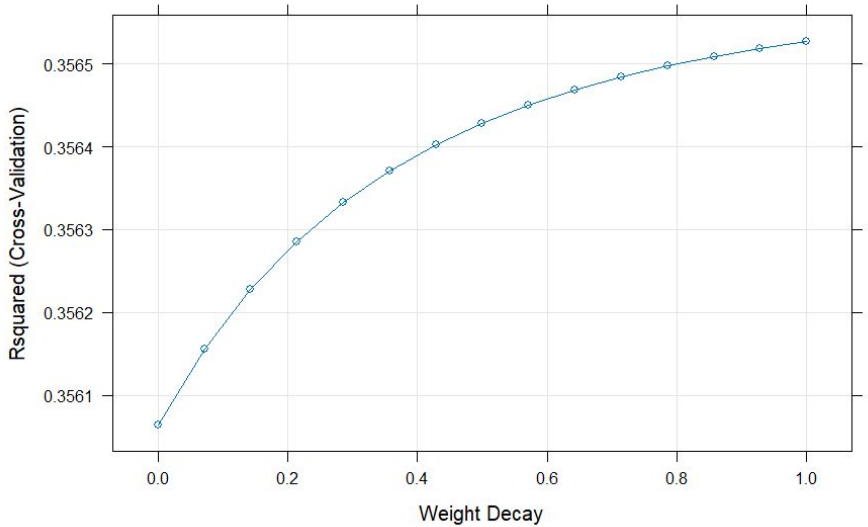
```
Ridge Regression

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4331, 4329, 4330, 4330, 4330, 4330, ...
Resampling results across tuning parameters:

  lambda      RMSE      Rsquared   MAE
  0.00000000  172.1014  0.3560645  76.84980
  0.07142857  172.0786  0.3561560  76.82779
  0.14285714  172.0599  0.3562281  76.80972
  0.21428571  172.0444  0.3562858  76.79475
  0.28571429  172.0314  0.3563325  76.78238
  0.35714286  172.0203  0.3563706  76.77231
  0.42857143  172.0108  0.3564021  76.76481
  0.50000000  172.0025  0.3564283  76.75837
  0.57142857  171.9954  0.3564502  76.75273
  0.64285714  171.9890  0.3564686  76.74793
  0.71428571  171.9835  0.3564843  76.74372
  0.78571429  171.9785  0.3564975  76.73995
  0.85714286  171.9741  0.3565089  76.73666
  0.92857143  171.9701  0.3565186  76.73374
  1.00000000  171.9665  0.3565269  76.73113

Rsquared was used to select the optimal model using the largest value.
The final value used for the model was lambda = 1.
```

Ridge Regression Tuning Parameter plot

Transformed data:



PCA:



Linear Models: Lasso Regression

Transformed data:

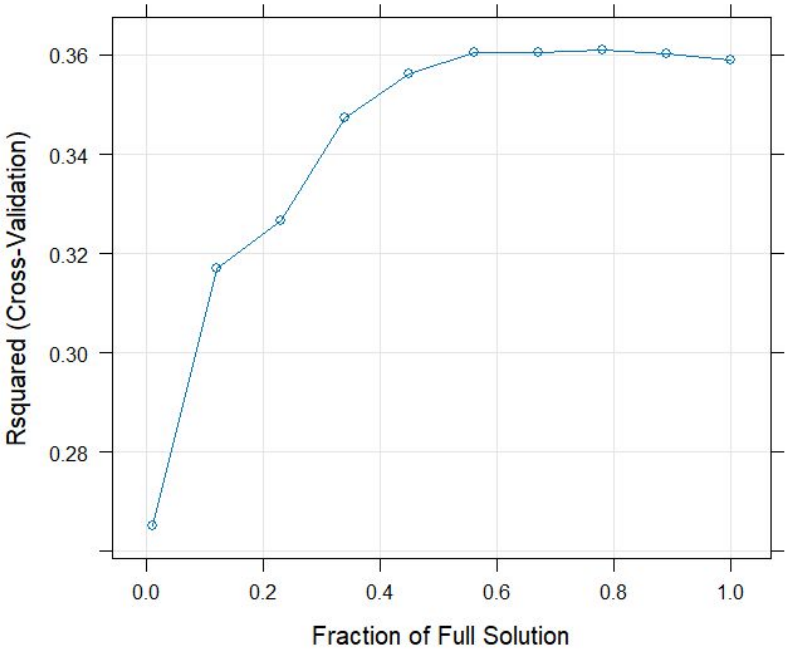
```
The lasso

4811 samples
  51 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4331, 4329, 4330, 4329, 4329, ...
Resampling results across tuning parameters:

  fraction  RMSE      Rsquared   MAE
  0.01      199.6517  0.3456396  104.38774
  0.12      179.5366  0.3456396   84.07439
  0.23      168.9596  0.3673373   74.03590
  0.34      165.8958  0.3852597   71.18518
  0.45      165.1026  0.3907636   70.64128
  0.56      164.9963  0.3918287   70.98729
  0.67      165.0595  0.3916703   71.53219
  0.78      165.2549  0.3906419   72.08336
  0.89      165.4845  0.3892388   72.66075
  1.00      165.6922  0.3880146   73.24782

Rsquared was used to select the optimal model using
 the largest value.
The final value used for the model was fraction = 0.56.
```

PCA:

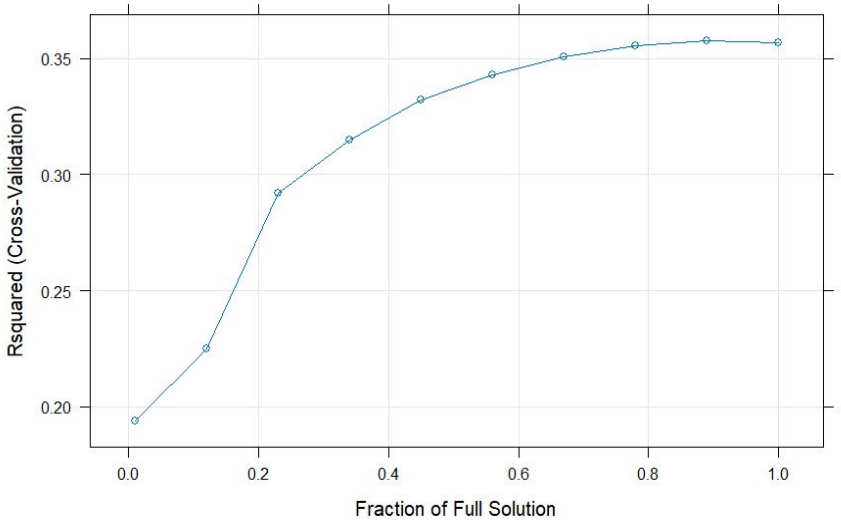
```
The lasso

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4331, 4331, 4330, 4329, 4329, 4329, ...
Resampling results across tuning parameters:

  fraction  RMSE      Rsquared   MAE
  0.01      208.7217  0.1940449  106.85337
  0.12      195.3033  0.2251446   92.01719
  0.23      186.6562  0.2920974   82.72732
  0.34      181.2658  0.3147239   77.69441
  0.45      178.2377  0.3320138   75.64801
  0.56      176.6050  0.3427803   74.64889
  0.67      175.5501  0.3506123   73.88594
  0.78      174.9530  0.3552882   73.55088
  0.89      174.7075  0.3574193   73.68791
  1.00      174.8753  0.3567233   74.21057

Rsquared was used to select the optimal model using the largest value.
The final value used for the model was fraction = 0.89.
```

Lasso Regression Tuning Parameter plot

Transformed data:



PCA:



Linear Models: PLS

Transformed data:

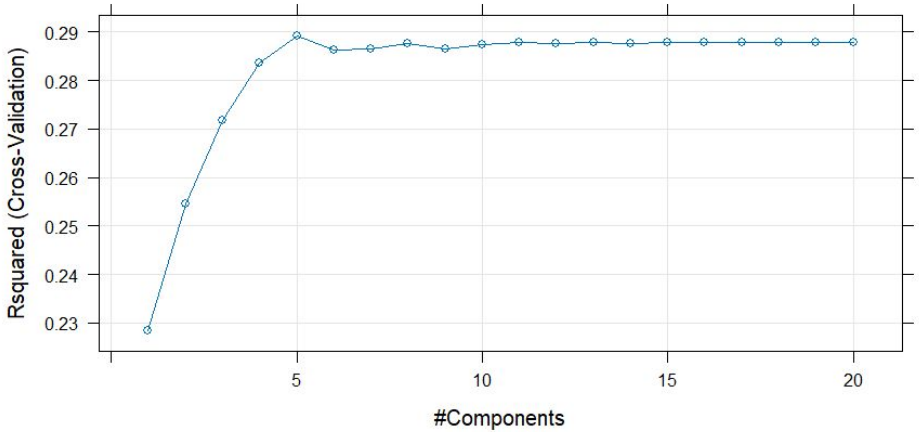
```
Partial Least Squares

4811 samples
  51 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4331, 4329, 4330, 4330, 4329, 4331, ...
Resampling results across tuning parameters:

  ncomp  RMSE      Rsquared   MAE
   1     194.7949  0.2283666  83.58557
   2     191.9194  0.2545692  85.48516
   3     189.9955  0.2716987  83.74821
   4     188.5977  0.2835071  82.91420
   5     187.9676  0.2892177  82.32951
   6     188.2852  0.2862569  82.69618
   7     188.2432  0.2865350  82.66068
   8     188.0839  0.2876428  82.81993
   9     188.2396  0.2864172  83.17400
  10     188.1054  0.2875118  83.01398
  11     188.0604  0.2878805  82.90831
  12     188.0785  0.2877033  82.85935
  13     188.0650  0.2877903  82.82968
  14     188.0718  0.2877366  82.82891
  15     188.0675  0.2877735  82.82570
  16     188.0619  0.2878258  82.81268
  17     188.0605  0.2878344  82.81262
  18     188.0622  0.2878177  82.81665
  19     188.0642  0.2878017  82.81743
  20     188.0650  0.2877946  82.81781

Rsquared was used to select the optimal model using the largest value.
The final value used for the model was ncomp = 5.
```

PCA:

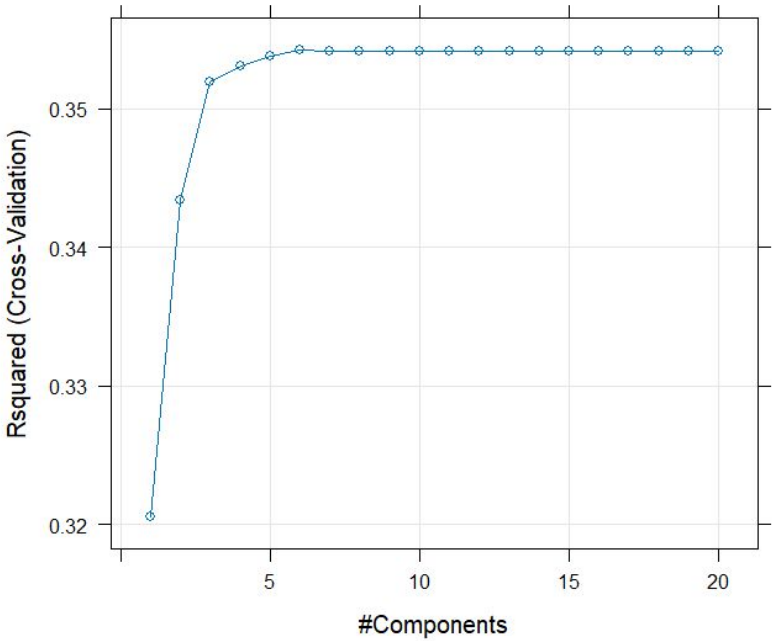
```
Partial Least Squares

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4331, 4329, 4328, 4330, 4329, 4330, ...
Resampling results across tuning parameters:

  ncomp  RMSE      Rsquared   MAE
   1     179.8467  0.3205646  76.40962
   2     177.3885  0.3433912  75.23586
   3     176.5344  0.3519133  74.39322
   4     176.3838  0.3530371  74.61072
   5     176.2872  0.3537751  74.52822
   6     176.2439  0.3541942  74.43035
   7     176.2471  0.3541654  74.44763
   8     176.2498  0.3541292  74.45621
   9     176.2495  0.3541328  74.45652
  10     176.2494  0.3541339  74.45569
  11     176.2494  0.3541338  74.45579
  12     176.2494  0.3541340  74.45577
  13     176.2494  0.3541340  74.45578
  14     176.2494  0.3541340  74.45578
  15     176.2494  0.3541340  74.45578
  16     176.2494  0.3541340  74.45578
  17     176.2494  0.3541340  74.45578
  18     176.2494  0.3541340  74.45578
  19     176.2494  0.3541340  74.45578
  20     176.2494  0.3541340  74.45578

Rsquared was used to select the optimal model using the largest value.
The final value used for the model was ncomp = 6.
```

PLS Tuning Parameter plot

Transformed data:



PCA:



Linear Models: ENet

Transformed data:

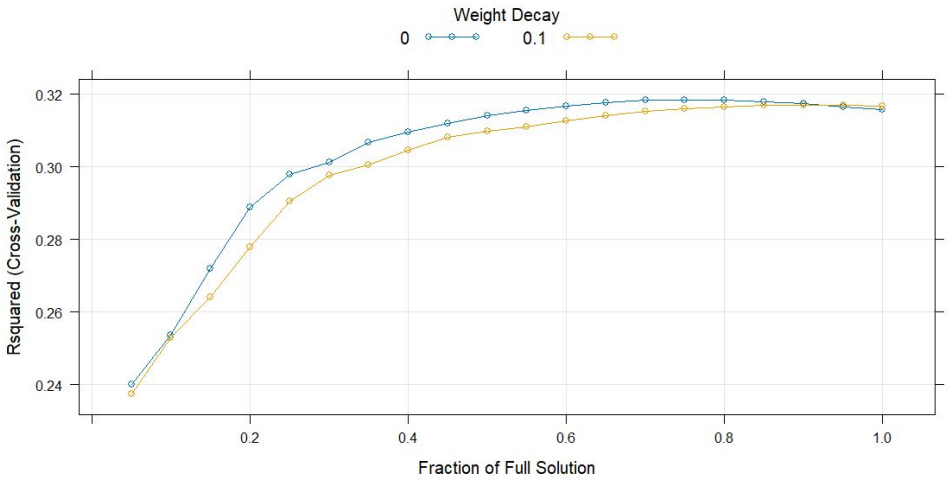
```
Elasticnet

4811 samples
  51 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4330, 4331, 4329, 4331, 4331, 4329, ...
Resampling results across tuning parameters:

  lambda  fraction  RMSE      Rsquared   MAE
  0.0     0.05      192.2618  0.3410622  95.92722
  0.0     0.10      183.4841  0.3410622  86.95372
  0.0     0.15      176.7331  0.3464340  80.13679
  0.0     0.20      171.7976  0.3573025  75.48324
  0.0     0.25      168.9218  0.3651376  73.12244
  0.0     0.30      167.2828  0.3761592  71.54832
  0.0     0.35      166.3608  0.3819952  70.99301
  0.0     0.40      165.7775  0.3861895  70.69930
  0.0     0.45      165.4315  0.3887542  70.57115
  0.0     0.50      165.2684  0.3899014  70.65292
  0.0     0.55      165.1666  0.3905928  70.84421
  0.0     0.60      165.1004  0.3911229  71.08346
  0.0     0.65      165.0800  0.3913314  71.34971
  0.0     0.70      165.0666  0.3915617  71.59605
  0.0     0.75      165.1165  0.3912932  71.86743
  0.0     0.80      165.1948  0.3907819  72.12847

Rsquared was used to select the optimal model using
 the largest value.
The final values used for the model were fraction = 0.85
 and lambda = 0.1.
```

PCA:

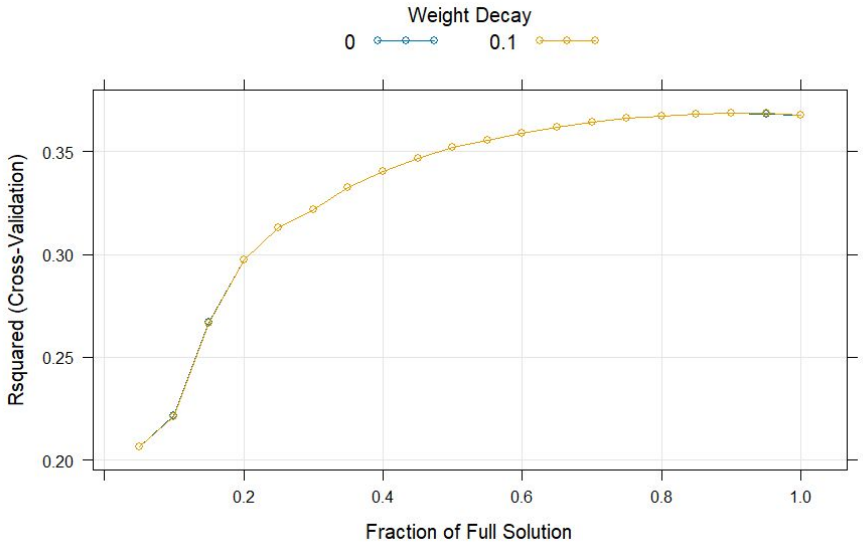
```
Elasticnet

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4331, 4330, 4331, 4330, 4329, ...
Resampling results across tuning parameters:

  lambda  fraction  RMSE      Rsquared   MAE
  0.0     0.05      199.6696  0.2065051  100.60275
  0.0     0.10      193.7273  0.2215214   94.28869
  0.0     0.15      188.9910  0.2668806   89.08764
  0.0     0.20      184.9017  0.2974271   84.85727
  0.0     0.25      181.4919  0.3131848   81.42745
  0.0     0.30      178.8580  0.3217276   79.03267
  0.0     0.35      176.8018  0.3324349   77.38859
  0.0     0.40      175.2397  0.3404262   76.34074
  0.0     0.45      174.0997  0.3469266   75.68899
  0.0     0.50      173.2376  0.3520072   75.19231
  0.0     0.55      172.6045  0.3557215   74.77335
  0.0     0.60      172.0788  0.3591462   74.40868
  0.0     0.65      171.6474  0.3620865   74.10265
  0.0     0.70      171.3052  0.3644775   73.86420

Rsquared was used to select the optimal model using the largest value.
The final values used for the model were fraction = 0.9 and lambda = 0.1.
```

ENet Tuning Parameter plot

Transformed data:



PCA:



Non-Linear Models: KNN

Transformed data:

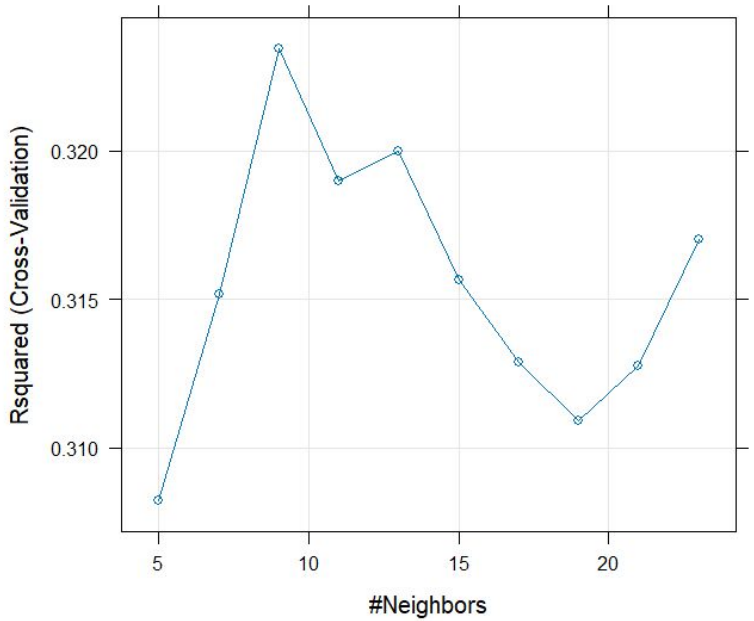
```
k-Nearest Neighbors

4811 samples
  51 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4330, 4331, 4330, 4329, 4331, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared   MAE
   5  181.0065  0.3082334  76.19479
   7  177.1163  0.3151714  75.73565
   9  175.0745  0.3234196  75.25901
  11  175.1944  0.3189893  75.58909
  13  174.9450  0.3199622  75.34060
  15  174.7724  0.3156638  75.55712
  17  174.8385  0.3128708  75.57241
  19  175.2416  0.3109281  75.79704
  21  174.9175  0.3127538  75.68922
  23  174.2256  0.3170163  75.57760

Rsquared was used to select the optimal model using
 the largest value.
The final value used for the model was k = 9.
```

PCA:

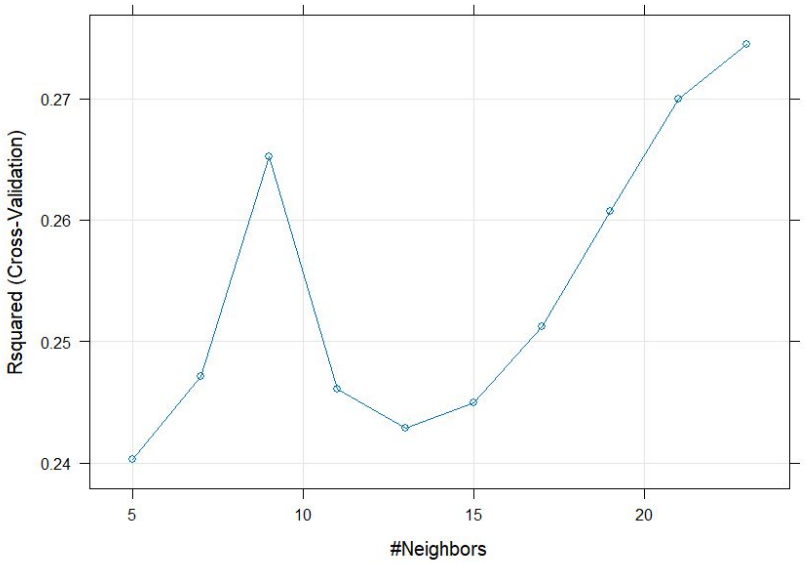
```
k-Nearest Neighbors

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4331, 4331, 4331, 4329, 4330, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared   MAE
   5  198.0070  0.2402544  81.81278
   7  190.5699  0.2471487  81.21572
   9  185.1319  0.2651805  80.87879
  11  186.8206  0.2460812  82.03979
  13  186.3777  0.2428922  83.09009
  15  184.7062  0.2449683  83.19663
  17  183.2005  0.2512166  82.83924
  19  181.6628  0.2606803  82.56905
  21  180.1771  0.2699906  82.37027
  23  179.2220  0.2744966  82.25399

Rsquared was used to select the optimal model using the largest value.
The final value used for the model was k = 23.
```

KNN Tuning Parameter plot

Transformed data:



PCA:



Non-Linear Models: MARS

Transformed data:

```
Multivariate Adaptive Regression Spline

4811 samples
  51 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4330, 4329, 4330, 4330, 4329, ...
Resampling results across tuning parameters:
```

degree	nprune	RMSE	Rsquared	MAE
1	1	217.2690	NaN	109.58554
1	2	192.8471	0.2437781	83.69281
1	3	196.6503	0.2210947	83.23550
1	4	195.5259	0.2496249	84.00332
1	5	191.1431	0.2824784	80.84573
1	6	189.5726	0.2912968	80.44371
1	7	187.7457	0.3061007	78.16424
1	8	186.7411	0.3137681	76.67263
1	9	184.1098	0.3328816	75.48911
1	10	184.8211	0.3277843	76.60373
1	11	182.7240	0.3454158	76.59033
1	12	183.0569	0.3427691	77.01511
1	13	183.0354	0.3435083	76.67144
1	14	185.0443	0.3316849	77.15032
1	15	183.8261	0.3398953	77.19078
1	16	184.7624	0.3302801	78.92505
1	17	184.9474	0.3315183	78.58877
1	18	184.8115	0.3314495	78.90059
1	19	185.0075	0.3303276	78.70617

PCA:

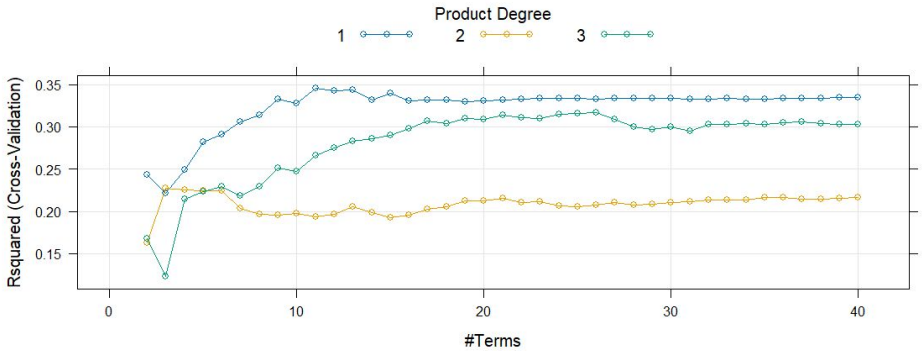
```
Multivariate Adaptive Regression Spline

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4330, 4331, 4329, 4332, 4329, ...
Resampling results across tuning parameters:
```

degree	nprune	RMSE	Rsquared	MAE
1	20	175.1740	0.3495510	77.61903
1	21	175.1549	0.3495878	77.61922
1	22	175.1007	0.3499884	77.62286
1	23	175.1443	0.3497622	77.74737
1	24	175.1752	0.3495049	77.77316
1	25	175.1282	0.3498896	77.72401
1	26	174.9943	0.3509834	77.62890
1	27	175.1154	0.3500017	77.68351
1	28	175.0988	0.3501338	77.69394
1	29	175.1296	0.3498893	77.73884
1	30	175.1296	0.3498893	77.73884
1	31	175.1296	0.3498893	77.73884
1	32	175.1296	0.3498893	77.73884
1	33	175.1296	0.3498893	77.73884
1	34	175.1296	0.3498893	77.73884
1	35	175.1296	0.3498893	77.73884

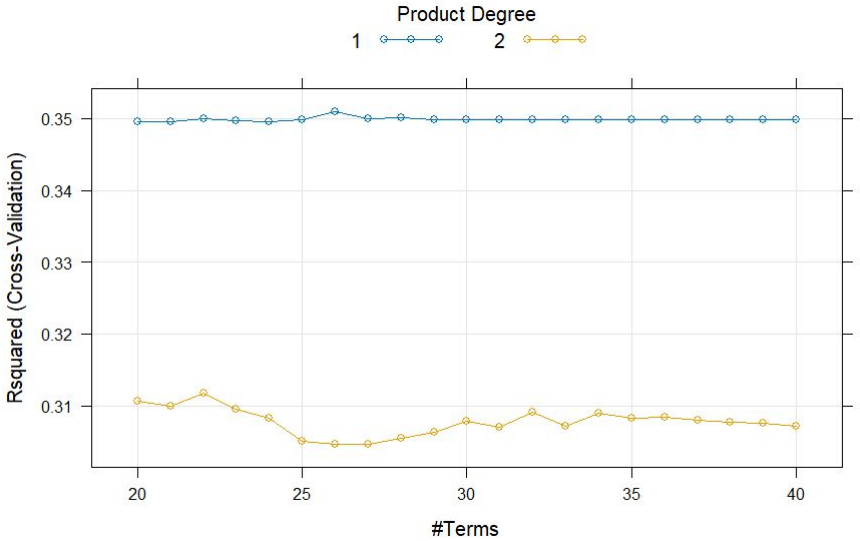
```
Rsquared was used to select the optimal model using the largest value.
The final values used for the model were nprune = 26 and degree = 1.
```

MARS Tuning Parameter plot

Transformed data:



PCA:



Non-Linear Models: Neural Networks

Transformed data:

```
Model Averaged Neural Network

4811 samples
  51 predictor

Pre-processing: centered (51), scaled (51)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4330, 4330, 4330, 4329, 4329, 4330, ...
Resampling results across tuning parameters:
```

decay	size	RMSE	Rsquared	MAE
0.0	1	186.0395	0.2764730	81.94301
0.0	2	177.7593	0.3450940	77.34338
0.0	3	179.4127	0.3335799	77.15948
0.0	4	181.3777	0.3260715	77.09401
0.0	5	187.5267	0.2984062	78.20051
0.0	6	184.6984	0.3226788	77.87447
0.0	7	179.9612	0.3503265	78.30160
0.0	8	187.2309	0.3188652	82.29842
0.0	9	195.4845	0.2801174	84.43461
0.0	10	179.9913	0.3427392	78.44338
0.1	1	182.1818	0.2981191	82.49044
0.1	2	179.7389	0.3292809	77.25977

PCA:

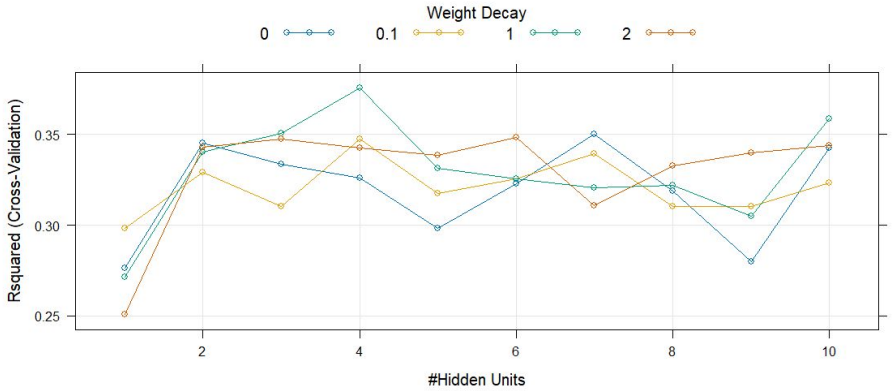
```
Neural Network

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4329, 4330, 4329, 4331, 4331, 4329, ...
Resampling results across tuning parameters:
```

decay	size	RMSE	Rsquared	MAE
0.0	1	188.5652	0.2256130	87.37561
0.0	2	185.4354	0.2682029	84.68665
0.0	3	184.3048	0.2881417	84.94503
0.0	4	284.7863	0.2200297	103.43085
0.0	5	200.9480	0.2579056	88.18413
0.0	6	200.9757	0.2551771	87.59168
0.0	7	240.2683	0.1954441	94.42212
0.0	8	203.4514	0.2694563	91.21548
0.0	9	202.7942	0.2599644	91.22612
0.0	10	202.2696	0.2292085	92.65087
0.0	11	190.0113	0.2738446	88.97045
0.0	12	185.0290	0.3014494	90.42464
0.0	13	184.8646	0.3077601	91.54292
0.0	14	184.8552	0.3069682	90.56659
0.0	15	188.4210	0.2957263	95.07917

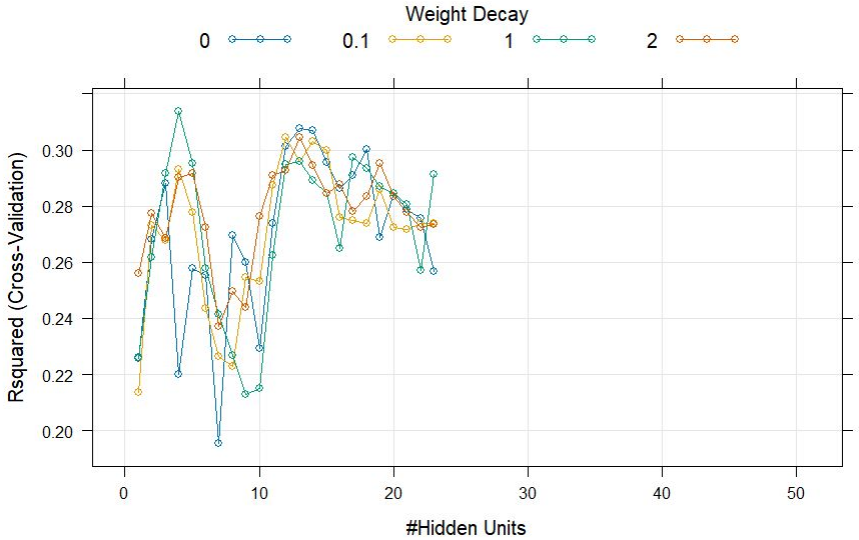
```
Rsquared was used to select the optimal model using the largest value.
The final values used for the model were size = 4 and decay = 1.
```

Neural Networks Tuning Parameter plot

Transformed data:



PCA:



Non-Linear Models: SVM

Transformed data:

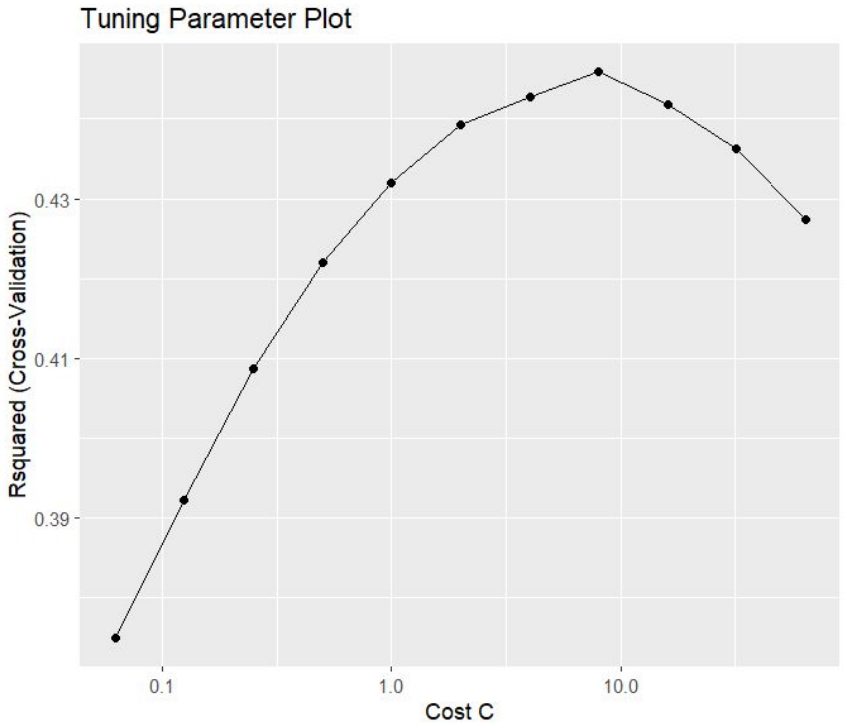
```
Support Vector Machines with Radial Basis Function Kernel

4811 samples
  51 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4330, 4329, 4331, 4330, 4331, 4329, ...
Resampling results across tuning parameters:

  C        RMSE      Rsquared   MAE
   0.0625  159.4851  0.4379685  63.88916
   0.1250  156.8420  0.4537712  62.37724
   0.2500  154.4191  0.4693711  60.90166
   0.5000  152.3252  0.4823874  59.78295
   1.0000  150.1896  0.4962392  59.10676
   2.0000  148.3024  0.5084063  58.86002
   4.0000  146.8068  0.5163006  59.02536
   8.0000  146.1581  0.5178973  59.86538
  16.0000  147.2573  0.5078275  62.01593
  32.0000  149.1987  0.4948780  65.45220
  64.0000  152.4181  0.4761200  69.53992

Tuning parameter 'sigma' was held constant at a value
 of 0.007971864
Rsquared was used to select the optimal model using
 the largest value.
The final values used for the model were sigma =
 0.007971864 and C = 8.
```

PCA:

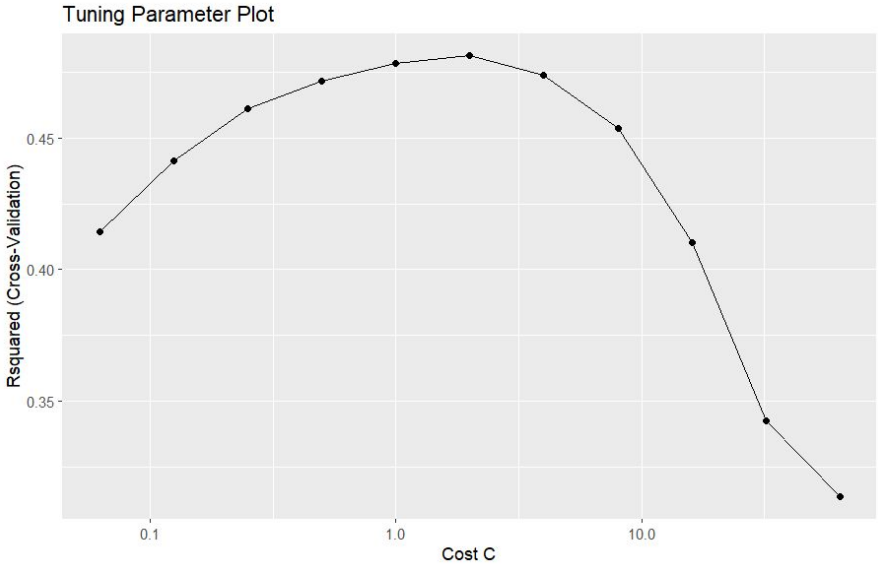
```
Support Vector Machines with Radial Basis Function Kernel

4811 samples
  40 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4331, 4330, 4331, 4330, 4330, 4329, ...
Resampling results across tuning parameters:

  C        RMSE      Rsquared   MAE
   0.0625  172.8283  0.4045249  67.12514
   0.1250  169.6624  0.4162057  65.25970
   0.2500  167.0145  0.4285970  63.79558
   0.5000  164.9803  0.4386853  62.76795
   1.0000  163.6617  0.4435465  62.24477
   2.0000  162.7905  0.4459634  62.38566
   4.0000  161.7112  0.4503961  62.57951
   8.0000  161.0693  0.4512730  63.55809
  16.0000  161.5282  0.4445107  65.71465
  32.0000  163.1093  0.4329039  69.16121
  64.0000  165.6077  0.4193104  73.34433

Tuning parameter 'sigma' was held constant at a value of 0.009628797
Rsquared was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.009628797 and C = 8.
```

SVM Tuning Parameter plot

Transformed data:



PCA:



Transformed Data Summary

Model	Best Tuning Parameter	Training RMSE	Testing RMSE	Training R2	Testing R2
OLR	Intercept = TRUE	169.46854402566	196.004784989907	0.355334196247485	0.273479466166533
Ridge	Lambda = 0.0714285714285714	175.546466999965	194.802516322503	0.378339186482991	0.280843748613416
Lasso	Fraction = 0.56	199.651664059443	193.394261076826	0.391828657809533	0.294732961848405
Enet	Fraction = 0.85 , Lambda = 0.1	193.301552359747	192.590891112119	0.392792809962208	0.297543204584997
KNN	k = 9	181.006511302703	195.909322945989	0.323419634412772	0.272419237545605
MARS	nPrune = 33 , Degree = 3	310.522334187129	228.476925779247	0.360801600629564	0.159016708333448
NN	size = 3 , decay = 1 , bag = FALSE	375.895630278195	255.661583852556	0.30359758360134	0.123893930632202
SVM	Sigma = 0.00797186442832724 , C = 8	159.485143781297	189.905026061827	0.517897310365258	0.323158595211138
PLS	Components = 5	169.120207757159	192.586411577671	0.391736188877329	0.298429011490971

PCA Summary

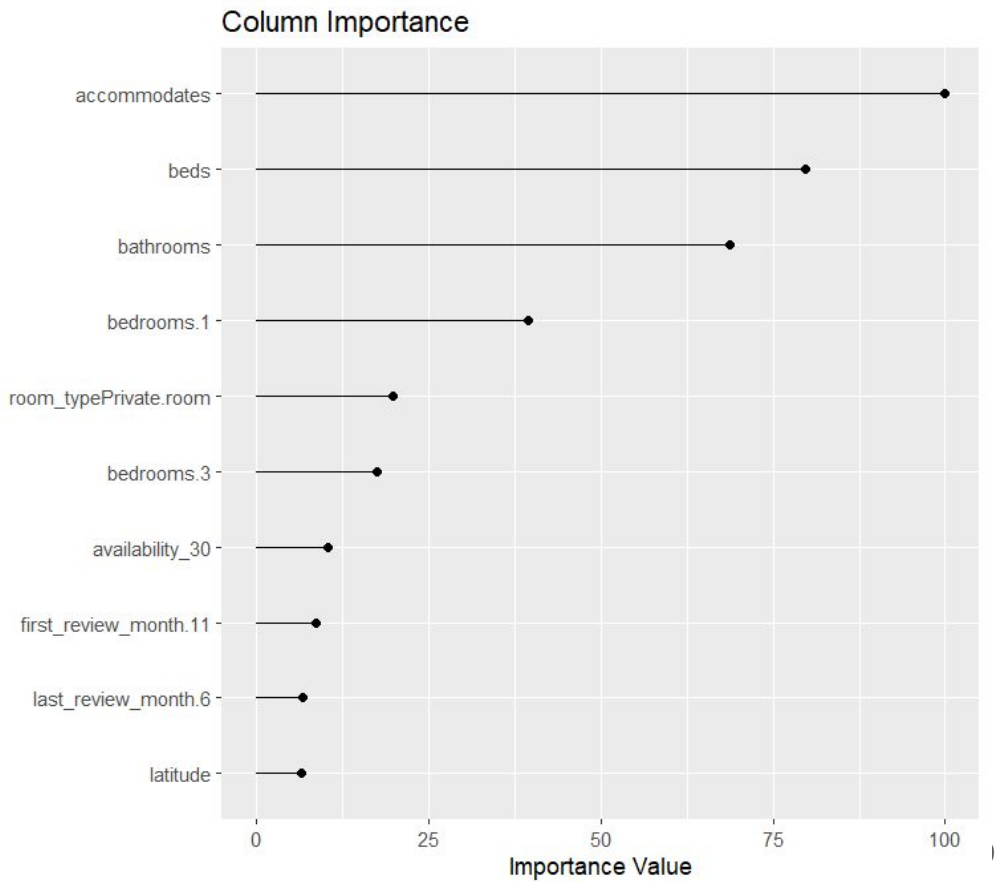
Model	Best Tuning Parameter	Training RMSE	Testing RMSE	Training R2	Testing R2
OLR	Intercept = TRUE	176.2767	174.9280	0.3340151	0.2751043
Ridge	Lambda = 1	180.0910	174.9524	0.3249482	0.2747824
PLS	Components = 6	168.7364	172.0820	0.3941537	0.2985590
Lasso	Fraction = 0.89	209.3823	171.8002	0.3609099	0.2999767
Enet	Fraction = 0.9 , Lambda = 0.1	204.2978	171.8075	0.3604448	0.2999492
KNN	k = 23	200.0803	238.7392	0.2842670	0.0000376
MARS	nPrune = 34 , Degree = 1	185.3496	172.3980	0.3483935	0.2953622
NN	size = 17 , decay = 2 , bag = FALSE	230.7485	179.1262	0.3145506	0.2632488
SVM	Sigma = 0.00962879682716566 , C = 8	172.8283	169.1500	0.4512730	0.3231621

Results

- Both the datasets did not perform well with the models, this could be due to underlying factors such as the dataset not having the necessary predictors which are required by the response.
- There were no significant differences in transformed data vs PCA, the transformed data slightly performed better.
- The top two models that performed well on training set are SVM with transformed data and SVM with PCA.

Important predictors for the best model

- The best model is **SVM** with Sigma = 0.0079718644283272 4 and C = 8 .



Thank you,
Let us know any questions you have!

Karthik Garimella, Sandeep Alfred

MA 5790 – Fall 2024 – Team 12