Statistical Learning Project

APPLYING STATISTICAL LEARNING TECHNIQUES TO ANALYZE MOTORSPORTS USING

Formula 1

AS AN EXAMPLE

November, 2022

Uladzislau Luksha
MSc Data Science and Economics, UNIMI

Presentation Plan

Dataset Exploratory Analysis

- Variables Description
- Data Relabeling
- Correlaition Plot

Unsupervised Learning

- PCA
- K-Means Clustering
- Hierarchical Clustering

Supervised Learning

- LASSO Regression
- Linear Regression
- Random Forest Regression

DATASET EXPLORATION

Formula 1 dataset

750 players

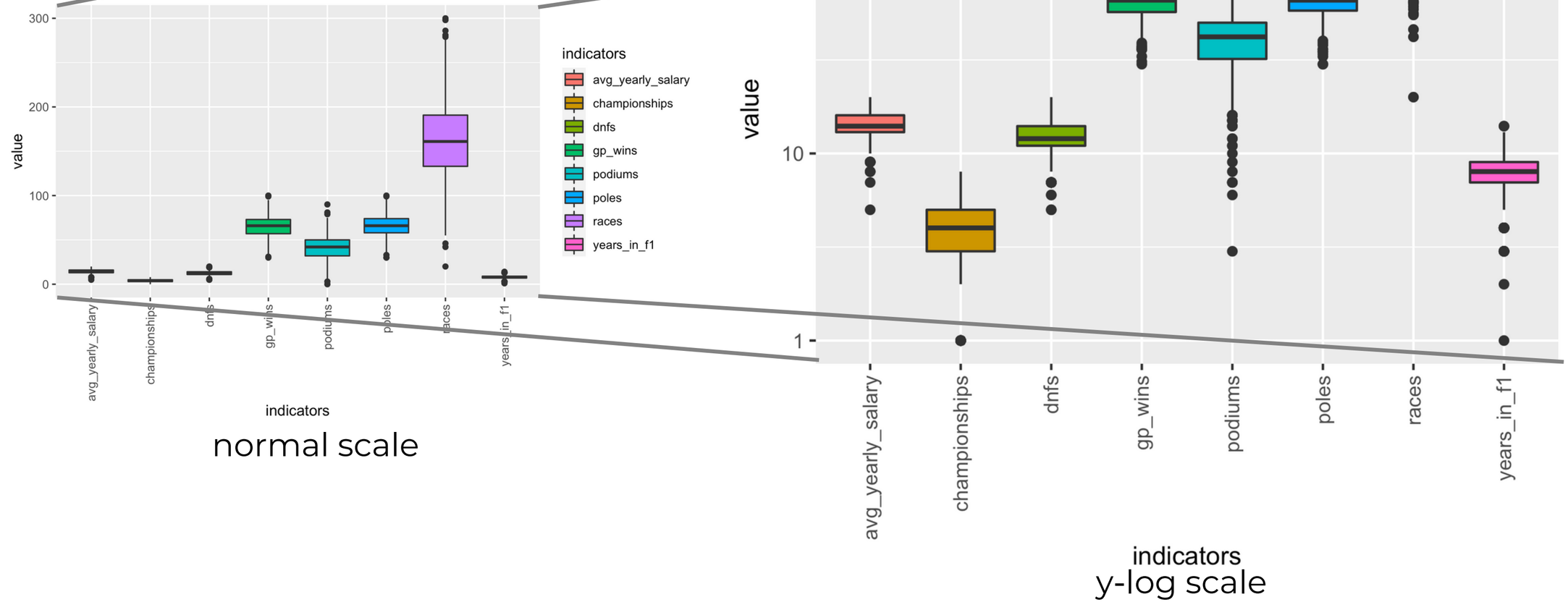
10 key performance indicators

- The f1 team that the driver drives for the major part of their career
- Number of GPs the driver has competed in
- Number of years a driver has spent in Formula 1
- Number of Championship Titles that the driver holds
- Number of races that the driver has won throughout his career
- Number of times a driver finished in the top-3 throughout his career
- Number of times a driver has started in the front of the
- Number of times the driver did not finish the race
- A driver's favorite circuit
- The average yearly salary that the driver earned throughout his career in millions of USD - ***Target Variable***

OUTLIERS

To detect **multivariate outliers** Isolation Forest was applied (threshold=0.595).
72 outliers removed for supervised techniques

univariate outliers detection with interquartile range



Data Relabeling

24 Circuit Names

grouped by geographical principal

‘APAC’ - Asia, Pasific Region, Australia

‘LATAM’ - South America + Mexico

'NORTH_AMERICA' - USA and Canada

‘MEAST_AFR’- Middle East and Africa

‘EUROPE’ - Europe

Data standardization (for unsupervised techniques),
One-hot encoding (for supervised techniques)
and zero variance check were performed

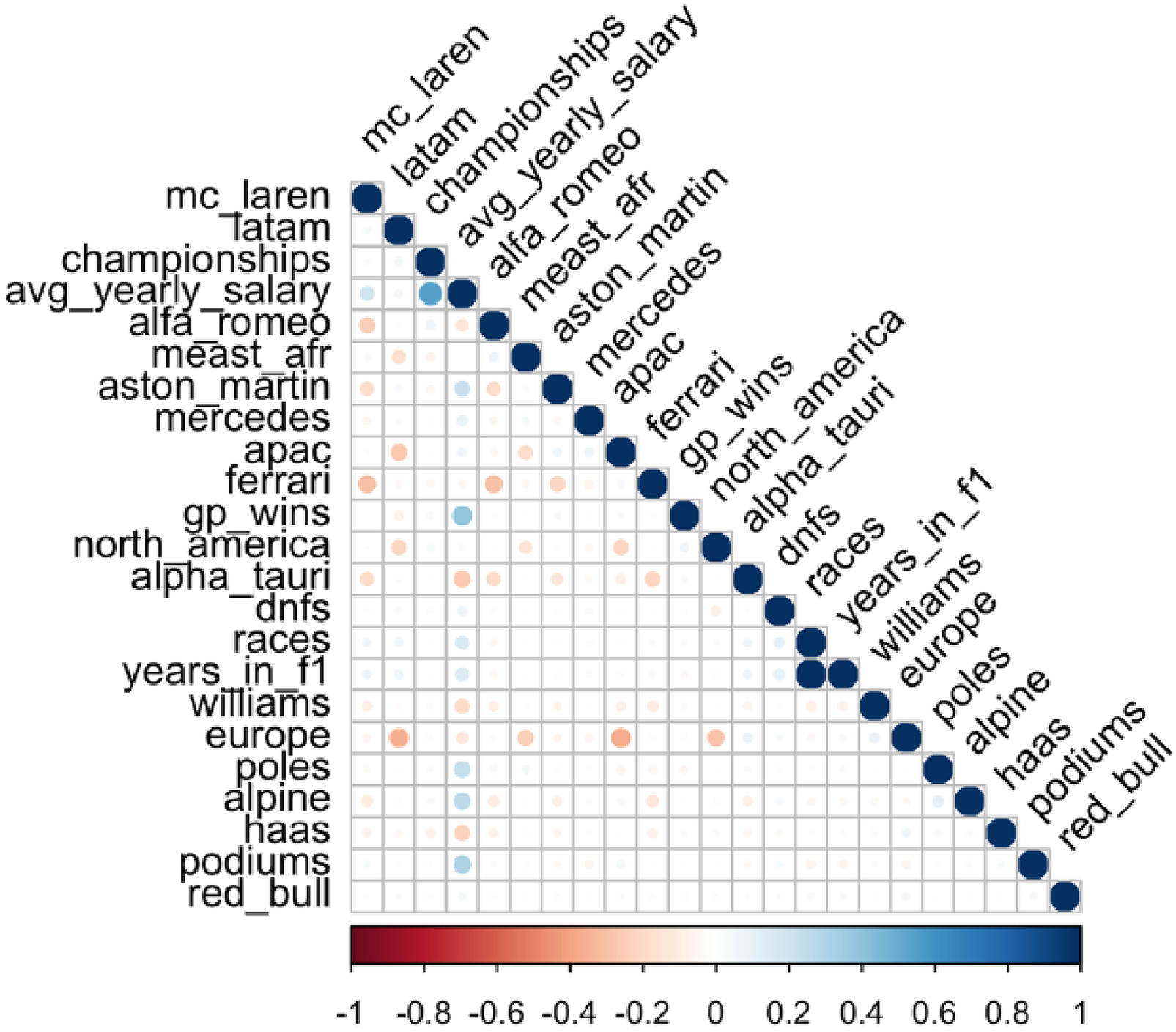
Correlation Plot

high correlation between years spent in Formula 1 and number of races for the player

No visible natural groups in the dataset

noticeable relationship between the dummy variables

problem of multicollinearity

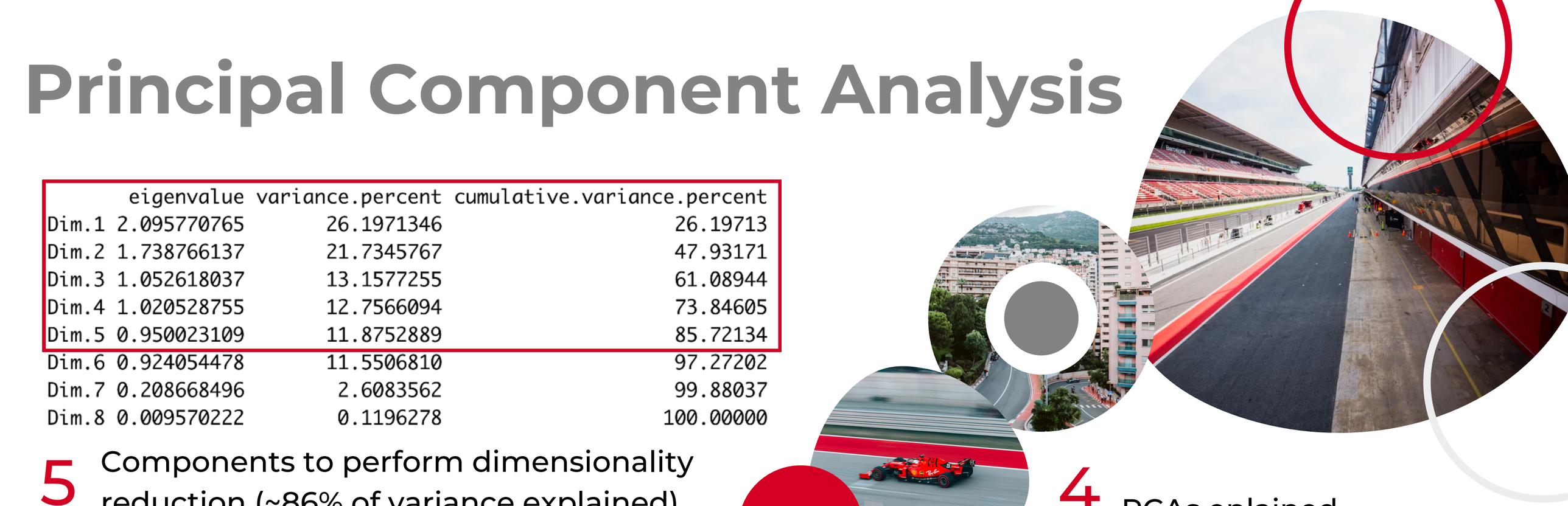


Unsupervised Learning

- PCA
- K-Means Clustering
- Hierarchical Clustering

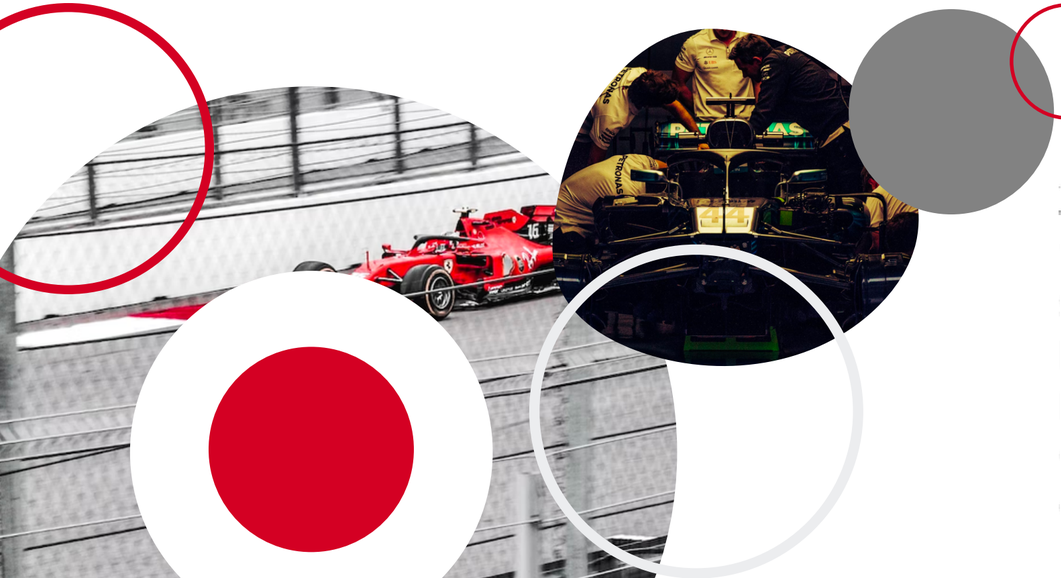
Principal Component Analysis

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.095770765	26.1971346	26.19713
Dim.2	1.738766137	21.7345767	47.93171
Dim.3	1.052618037	13.1577255	61.08944
Dim.4	1.020528755	12.7566094	73.84605
Dim.5	0.950023109	11.8752889	85.72134
Dim.6	0.924054478	11.5506810	97.27202
Dim.7	0.208668496	2.6083562	99.88037
Dim.8	0.009570222	0.1196278	100.00000



5 Components to perform dimensionality reduction (~86% of variance explained)

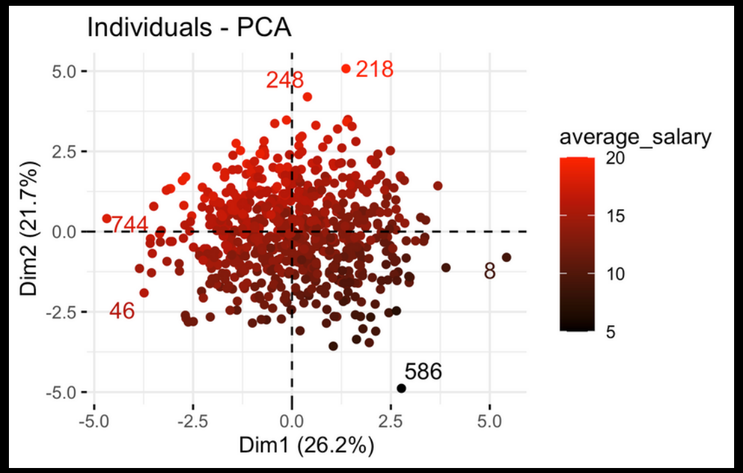
4 PCAs eplained

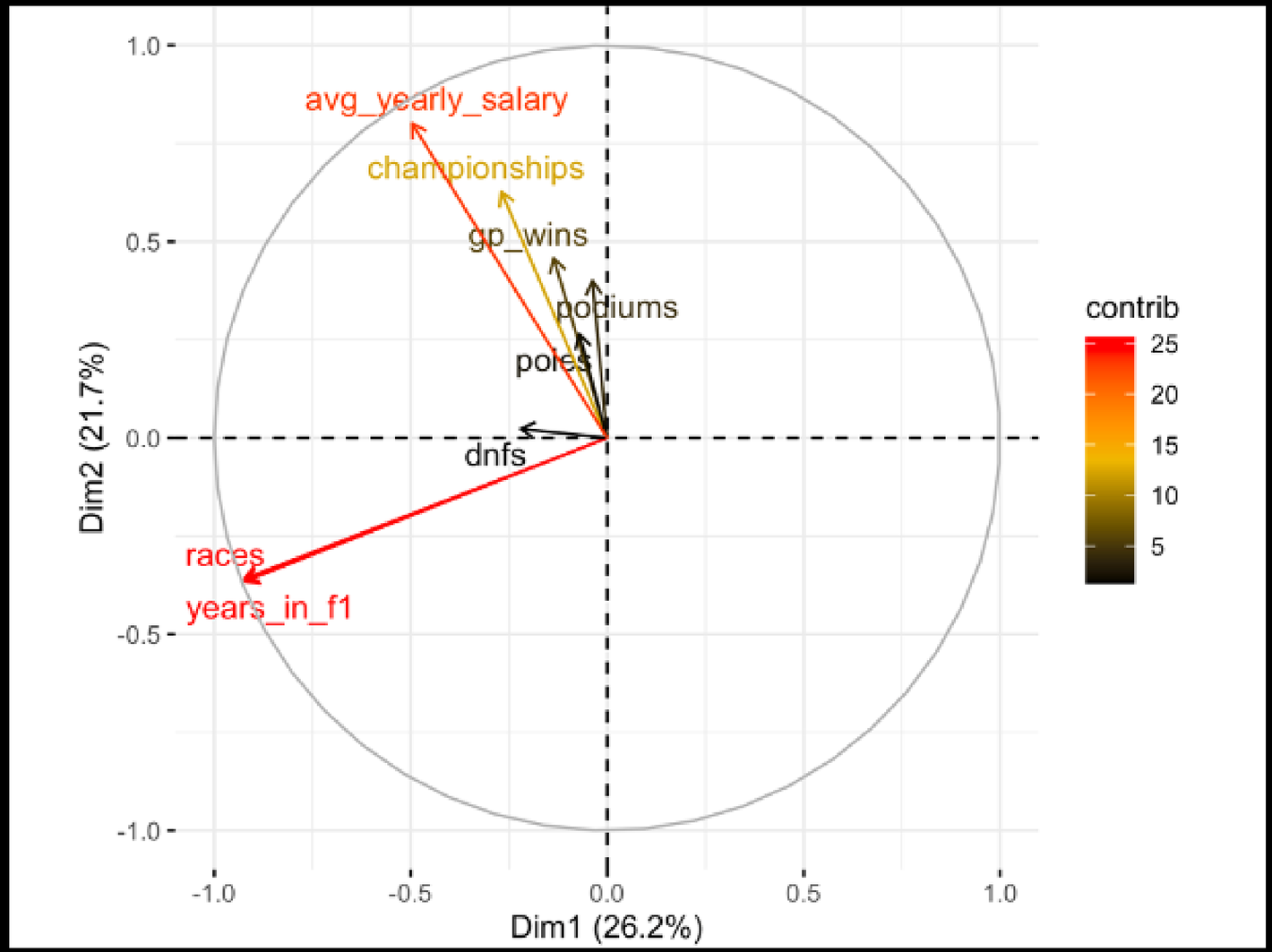


	PC1	PC2	PC3	PC4
races	-0.63819552	-0.27225760	-0.01559759	-0.02006993
years_in_f1	-0.63664002	-0.27716338	-0.01338186	-0.01476176
championships	-0.18613696	0.47581035	-0.11503518	-0.10974813
gp_wins	-0.09426305	0.34700509	0.55383797	-0.44505499
podiums	-0.02595256	0.30353922	0.08453401	0.71948754
poles	-0.04887155	0.20036694	-0.78930968	-0.07363667
dnfs	-0.15195814	0.01753696	0.22065889	0.51528613
avg_yearly_salary	-0.34308925	0.60799484	-0.02760335	-0.02561675

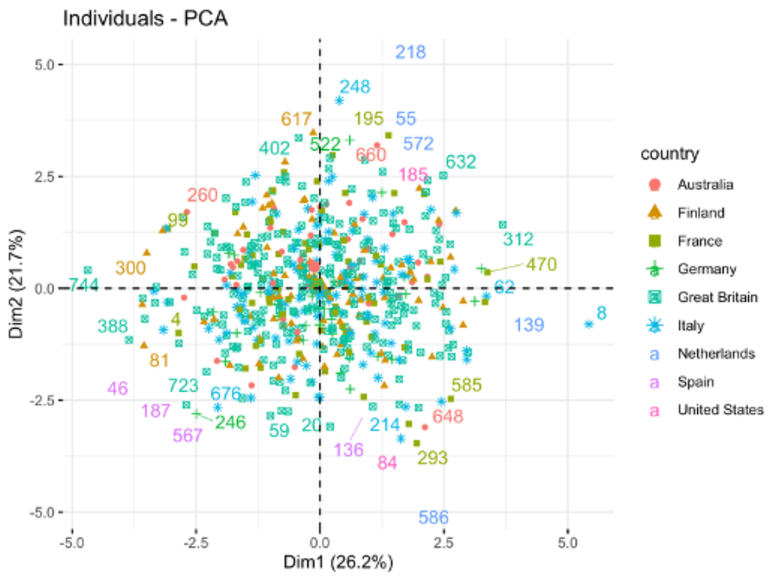
Principal Component Analysis

- Numeric variables plotted against PC1 and PC2
- Color - contributions of the variables to the PCs (loadings)
- Distance between variables and the origin - the quality of the representation by PCs
- average yearly salary, number of championship titles and gp_wins are positively correlated



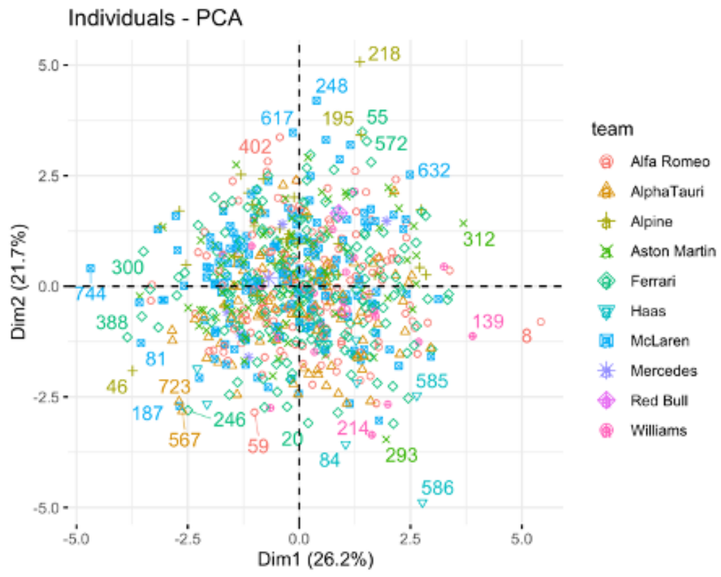


PCA colored by categorical variables



Country

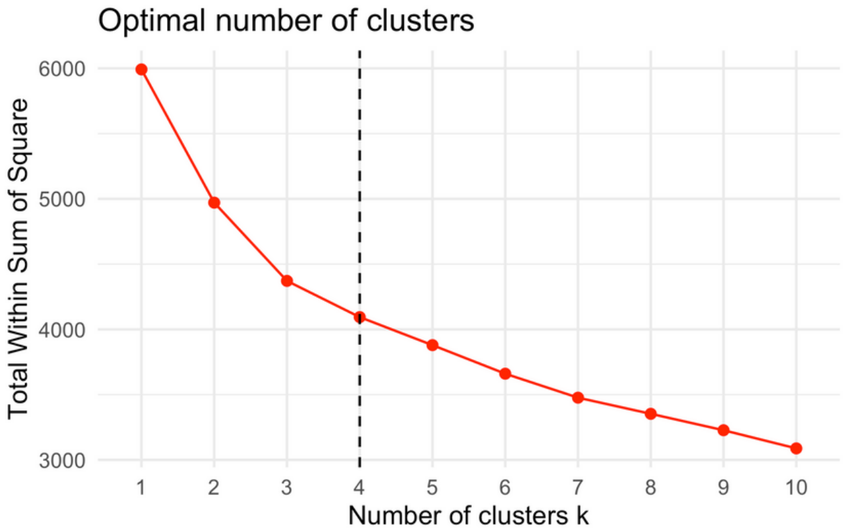
Favorite circuit



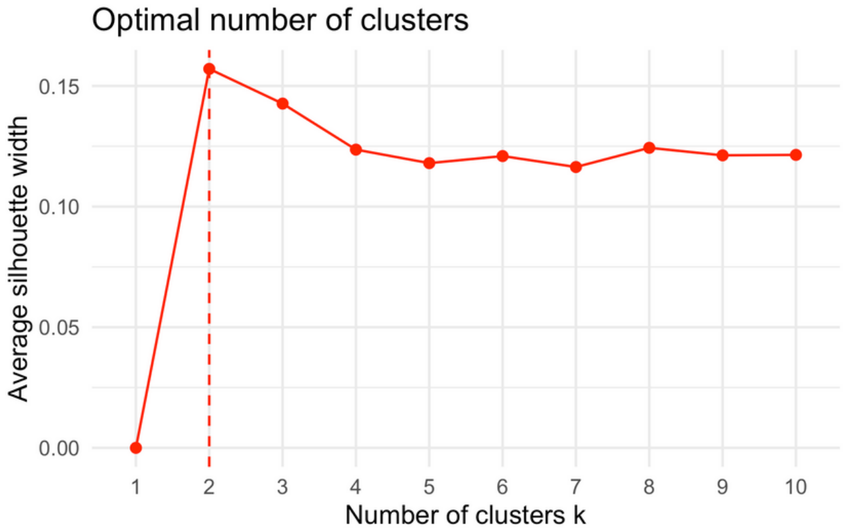
Team

K-Means Clustering

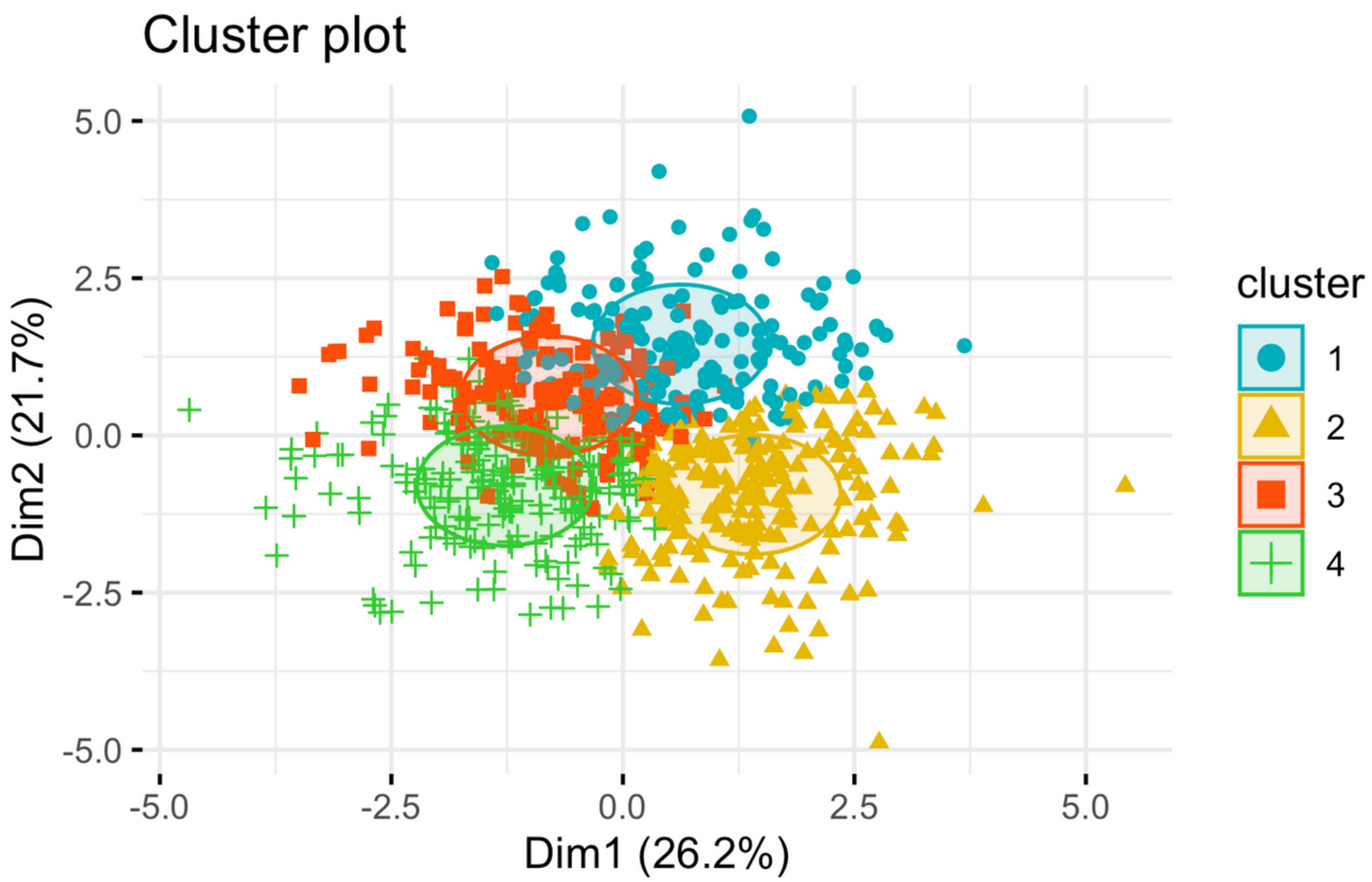
Optimal number of clusters K=4.
Eucleadian distance.



Elbow method



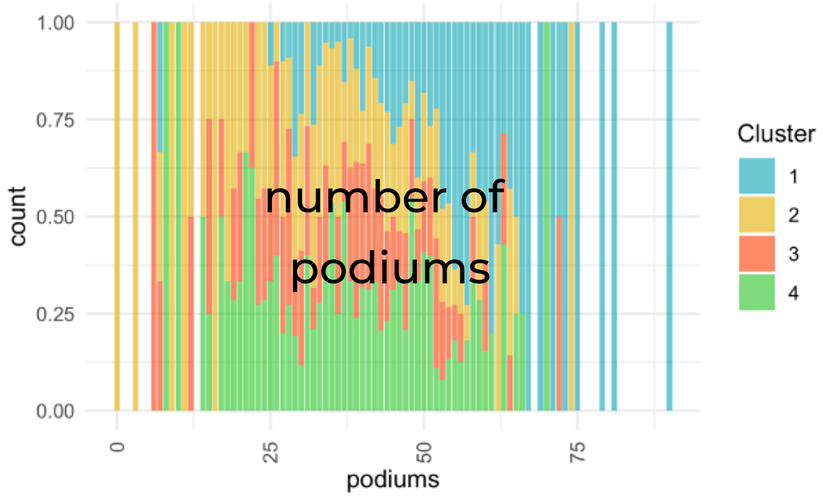
Silhouette method

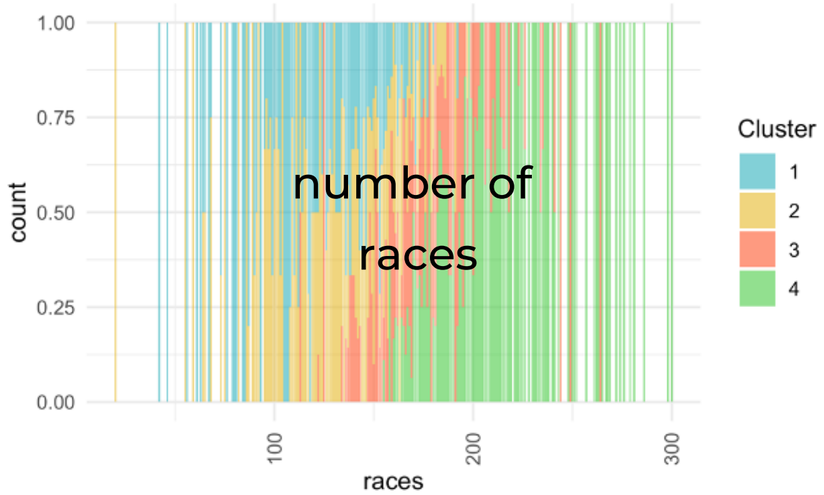


Cluster Plot of F1 Players

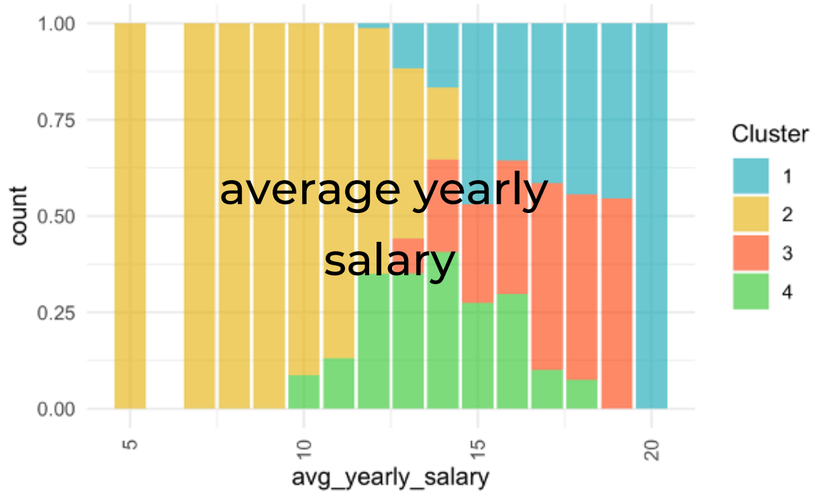
K-Means Clustering

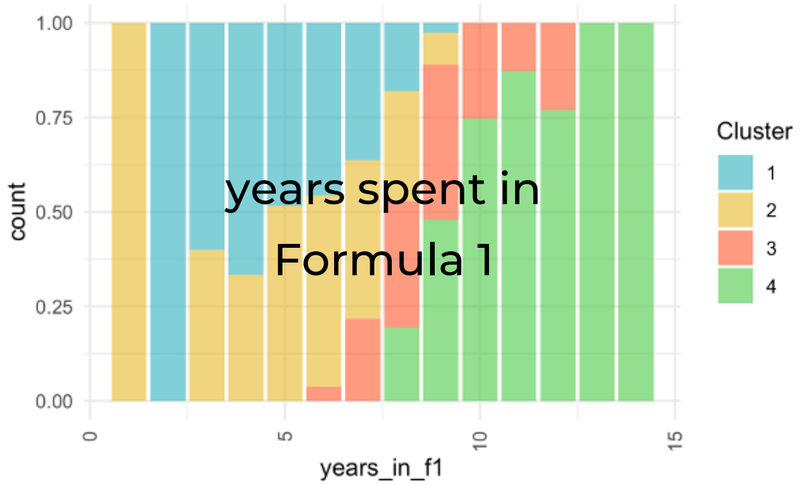
Optimal number of clusters K=4.
Eucleadian distance.





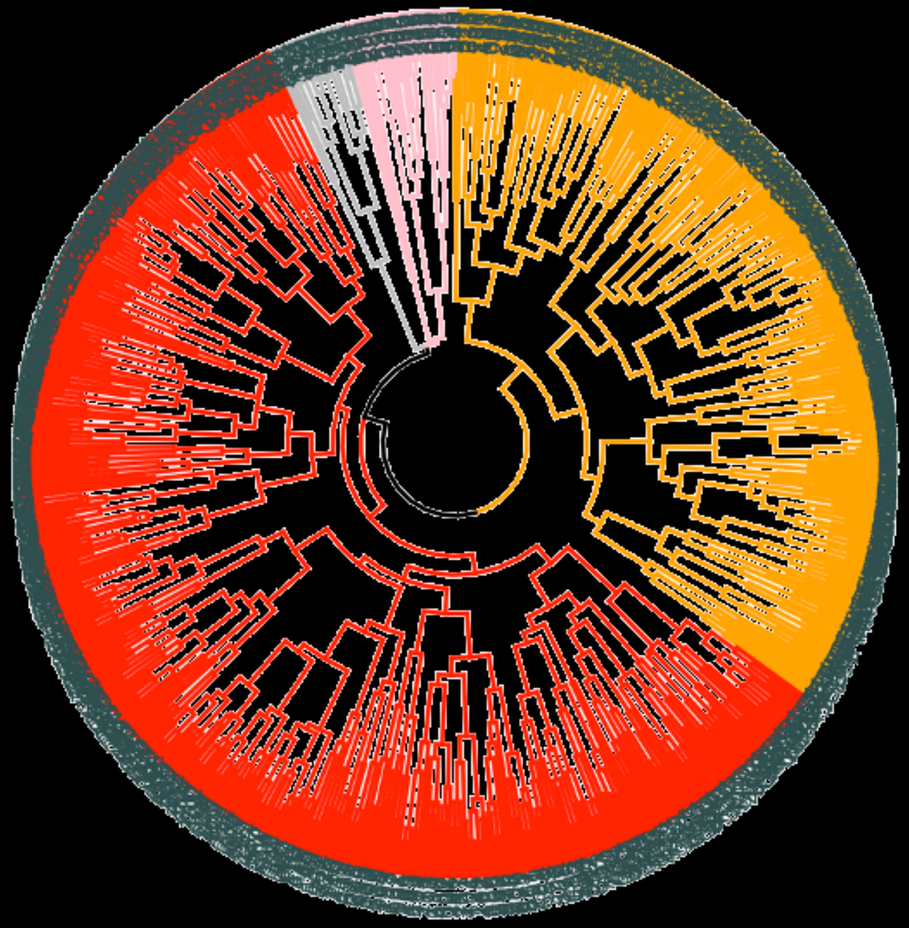
Cluster associations



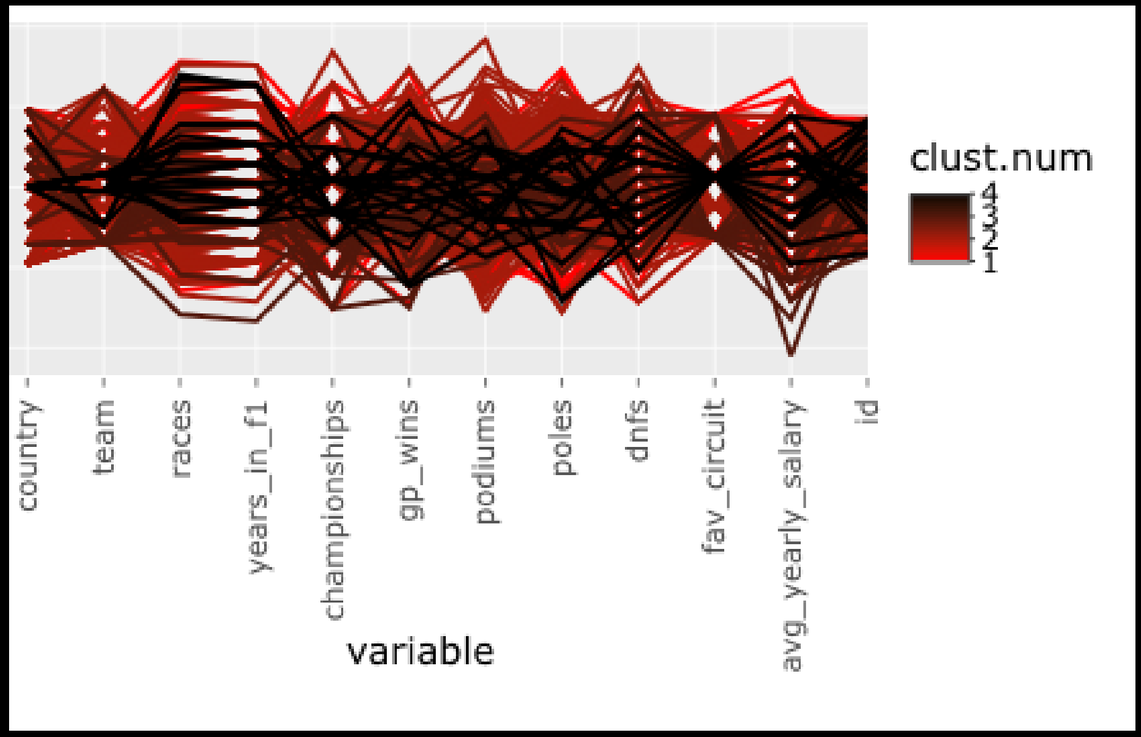


Hierarchical Clustering

- Categorical variables included
- Gower's dissimilarity measure
- Ward linkage applied



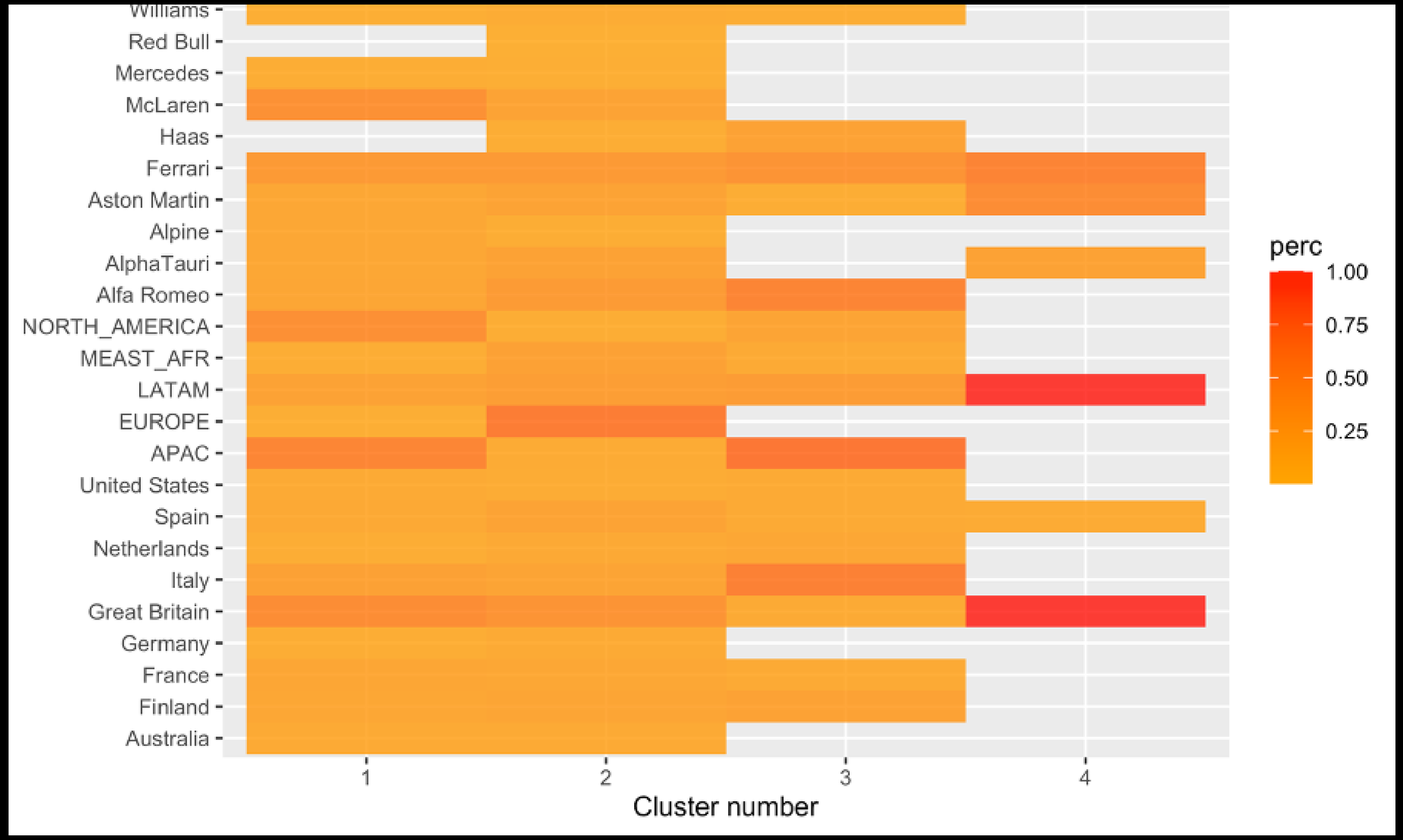
Dendrogram of Hierarchical Clustering (radial form)



Distribution of characteristics across clusters

Hierarchical Clustering

- Categorical variables included
- Gower's dissimilarity measure
- Ward linkage applied



Distribution of characteristics across clusters

Supervised Learning

data inference

- LASSO Regression
- Linear Regression

data prediction

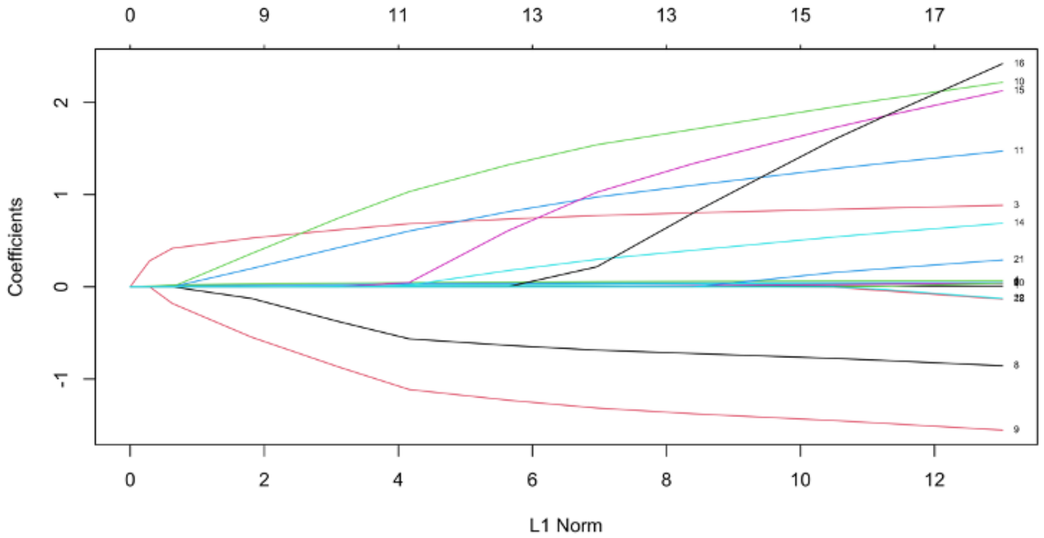
- Random Forest Regression

The average yearly salary that the driver earned throughout his career in millions of USD - Target Variable

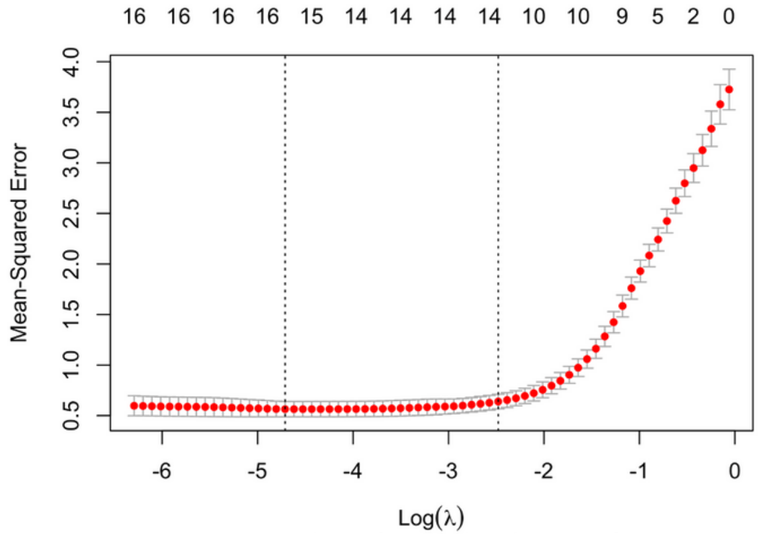
LASSO Regression

R-squared of the model is 0.872358

λ = 0.009



Regularization term graph for LASSO

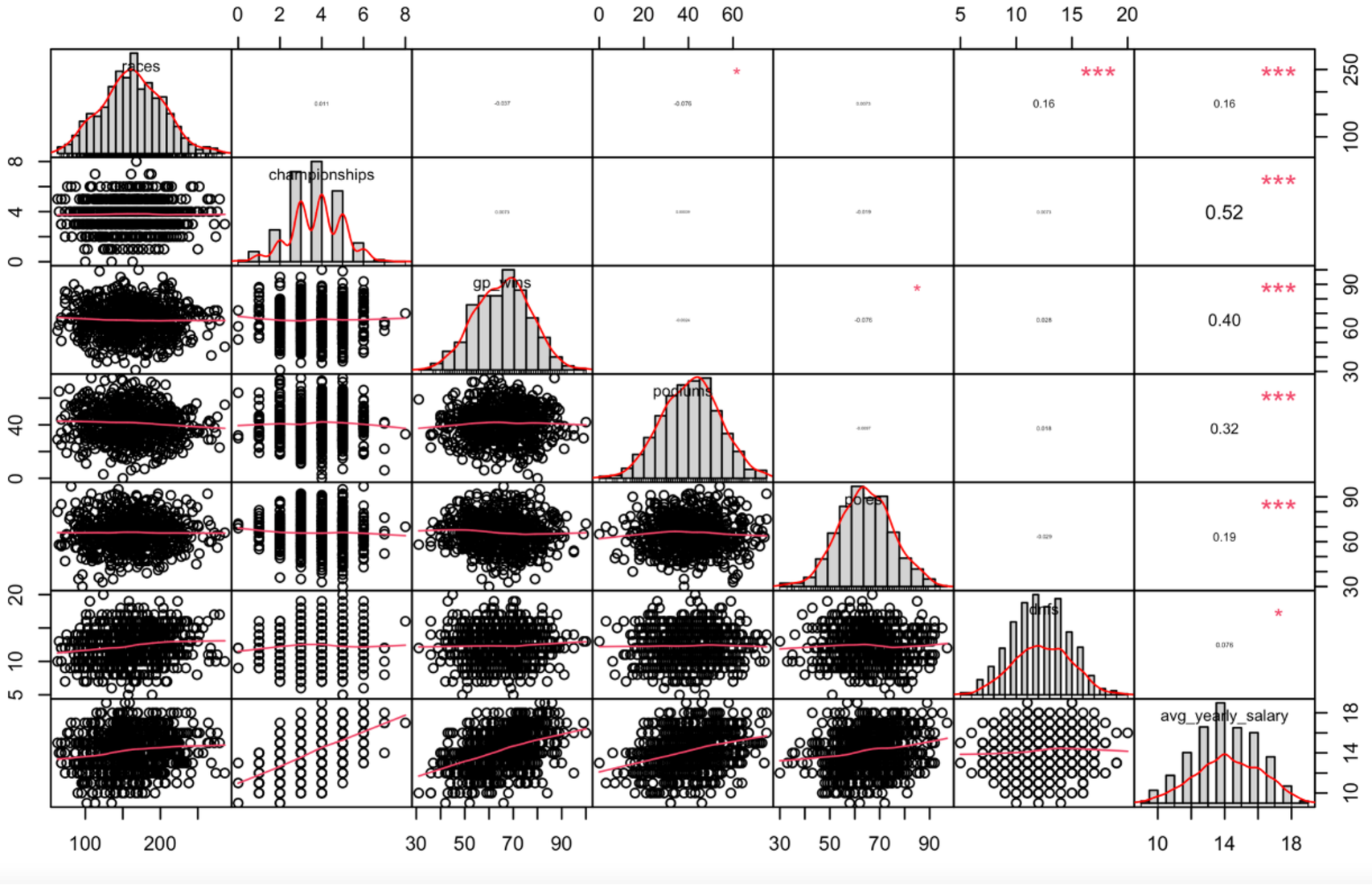


MSE by lambda value

(Intercept)	0.241404044	alpha_tauri	-2.284194146
races	0.008996112	alpine	1.303760819
championships	0.880457811	aston_martin	0.732672462
gp_wins	0.069856537	ferrari	-0.716627559
podiums	0.049499427	mc_laren	.
poles	0.045394672	apac	.
dnfs	0.019194151	europe	.
alfa_romeo	-1.485367367	latam	0.119428046
		meast_afr	0.292508845

Model Coefficients provided by LASSO regression

Linear Regression



Distribution of the variables used for the linear regression model and their correlation plots

Linear Regression (all features)

```
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)    1.3902735  0.3263287   4.260              0.0000234 ***
races          0.0092228  0.0006342  14.543 < 0.0000000000000002 ***
championships  0.8968855  0.0211402  42.426 < 0.0000000000000002 ***
gp_wins        0.0700381  0.0022668  30.897 < 0.0000000000000002 ***
podiums        0.0500767  0.0019974  25.070 < 0.0000000000000002 ***
poles          0.0464648  0.0023123  20.095 < 0.0000000000000002 ***
dnfs           0.0209095  0.0101844   2.053              0.040455 *
alfa_romeo    -2.8886647  0.1422083 -20.313 < 0.0000000000000002 ***
alpha_tauri   -3.6907353  0.1490692 -24.759 < 0.0000000000000002 ***
aston_martin  -0.6187678  0.1498373  -4.130              0.0000410 ***
ferrari       -2.1073655  0.1394532 -15.112 < 0.0000000000000002 ***
mc_laren      -1.3790648  0.1426576  -9.667 < 0.0000000000000002 ***
latam          0.1331266  0.0640691   2.078              0.038107 *
meast_afr      0.3135398  0.0861949   3.638              0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6651 on 662 degrees of freedom
Multiple R-squared:  0.888,     Adjusted R-squared:  0.8858
F-statistic: 403.8 on 13 and 662 DF,  p-value: < 0.00000000000000022
```

Linear Regression results (all features)

```
> vif(step.model)
       races championships       gp_wins       podiums         poles
    1.046358      1.018898      1.020250      1.009952      1.018895
        dnfs    alfa_romeo   alpha_tauri  aston_martin       ferrari
    1.039902      5.181650      3.808501      3.734090      6.063249
    mc_laren         latam     meast_afr
    5.161352      1.051756      1.054089
> sqrt(vif(mod)) > 2 # MULTICOLLINEARITY!
       races championships       gp_wins       podiums         poles
       FALSE         FALSE         FALSE         FALSE         FALSE
        dnfs    alfa_romeo   alpha_tauri        alpine  aston_martin
       FALSE          TRUE          TRUE          TRUE          TRUE
     ferrari      mc_laren          apac        europe         latam
        TRUE          TRUE         FALSE         FALSE         FALSE
   meast_afr
       FALSE
```

Variance inflation factor (all features)

Multicollinearity problem!

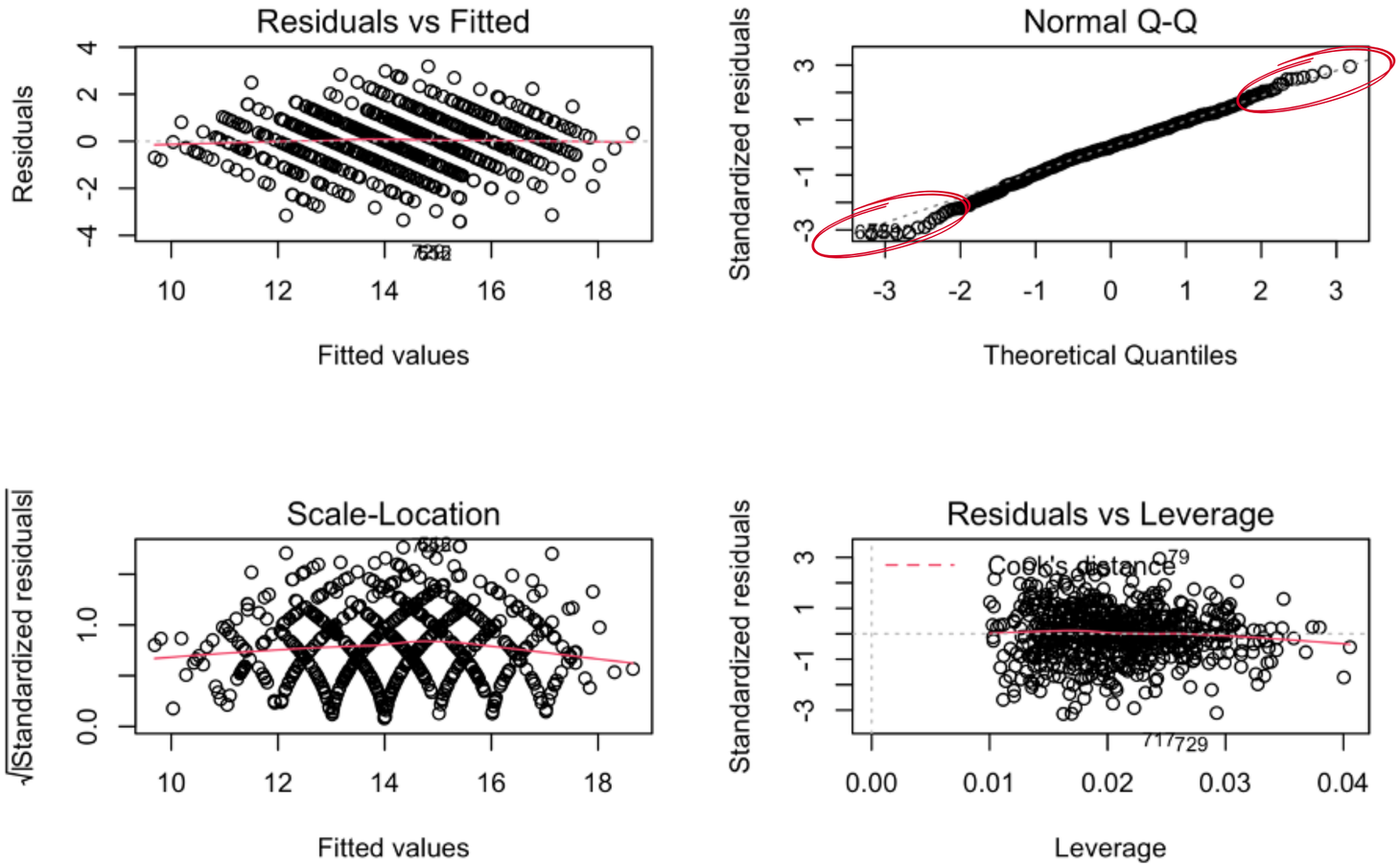
Linear Regression (adjusted features)

```
Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  -0.185671   0.456199  -0.407              0.68414
races         0.008281   0.001026   8.072  0.00000000000000326 ***
championships 0.873234   0.034685  25.176 < 0.0000000000000002 ***
gp_wins       0.074352   0.003715  20.012 < 0.0000000000000002 ***
podiums       0.050028   0.003288  15.215 < 0.0000000000000002 ***
poles         0.043490   0.003792  11.470 < 0.0000000000000002 ***
alfa_romeo   -0.988714   0.120822  -8.183  0.00000000000000141 ***
ferrari      -0.232052   0.110915  -2.092              0.03680 *
mc_laren      0.501900   0.120567   4.163  0.00003557468554005 ***
apac          0.293735   0.109846   2.674              0.00768 **
latam         0.301961   0.110586   2.731              0.00649 **
meast_afr     0.379608   0.145475   2.609              0.00927 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.092 on 664 degrees of freedom
Multiple R-squared:  0.697,     Adjusted R-squared:  0.692
F-statistic: 138.9 on 11 and 664 DF,  p-value: < 0.00000000000000022
```

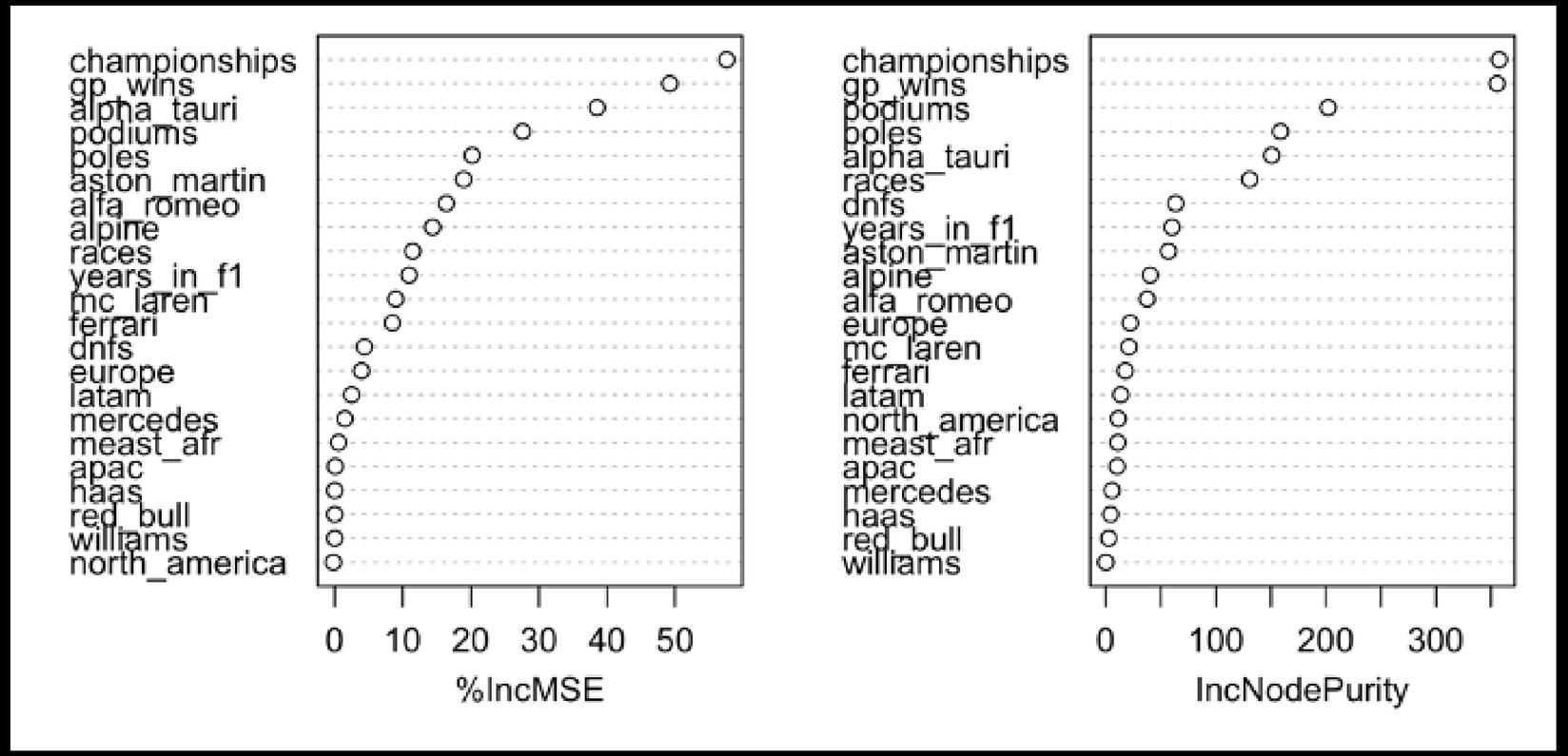
Linear Regression results (adjusted features)

No multicollinearity problem



Residuals versus fitted values, Scale-Location, residuals versus leverage, normal Q-Q plots for the adjusted linear regression model

Random Forest Regression



Importance of the indicators in Random Forest

- Train/Test split (p = 0.7)
- Bootstrap resampling
- 500 trees
- number of variables tried at each split = 5)

Variance explained

70.46%

train set

63.33%

test set

Low power of prediction

Conclusions

One of the most influential factor is the number of championship titles

K-Means is very sensitive to outliers, and Hierarchical Clustering does not allow to visualize its results even for relatively small datasets

K-means and Hierarchical Clustering are consistent and show that the dataset does not form any natural clusters

we encountered various statistical challenges: unbalanced data, the problem of competent management of outliers, multicollinearity of data, non-normal distribution of residuals.

Random Forest Regression showed rather mediocre results which can be caused by both the size of the dataset used and insufficient data quality

Techniques to improve the modelling:

- Relevant features selection
- Outliers removal
- Augmenting the dataset with synthetic samples
- Adding information from other sources.

Formula 1™

References

- f1-salary-prediction. (2022, October 13). Kaggle. https://www.kaggle.com/code/niramay/f1-salary-prediction/data
- GitHub - toUpperCase78/formula1-datasets: Datasets for Formula 1 World Championship. (n.d.). GitHub. Retrieved October 22, 2022, from https://github.com/toUpperCase78/formula1-datasets
- Staff, S. (2022, October 23). How much are F1 drivers paid? 2022 salaries revealed. The Independent. https://www.independent.co.uk/f1/how-much-drivers-paid-2022-salary-b2211780.html
- Khandelwal, A. (2021, September 6). Q-Q plot – Ensure Your ML Model is Based on the Right Distribution. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/q-q-plot-ensure-your-ml-model-is-based-on-the-right-distributions/
- Singh, D. (2019, November 12). Linear, Lasso, and Ridge Regression with R. Pluralsight. http://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r
- Songhao, Wu (2020, May 19) Multicollinearity in Regression. Retrieved October 24, 2022, from https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea#:~:text=Multicollinearity%20happens%20when%20independent%20variables,variables%20into%20the%20regression%20model
- Dummy Variable Trap. LearnDataSci. (n.d.). Retrieved October 26, 2022, from https://www.learndatasci.com/glossary/dummy-variable-trap/
- Reusova, A (2018, April 1) Hierarchical Clustering on Categorical Data in R. Retrieved October 24, 2022, from https://towardsdatascience.com/hierarchical-clustering-on-categorical-data-in-r-a27e578f2995
- F1 - The Official Home of Formula 1® Racing. (2022, March 18). Formula 1® - the Official F1® Website. https://www.formula1.com/en.html
- Le Roux; B. and H. Rouanet (2004). Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. Dordrecht. Kluwer: p.180.
- MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.
- Nielsen, Frank (2016). "8. Hierarchical Clustering". Introduction to HPC with MPI for Data Science. Springer. pp. 195–211. ISBN 978-3-319-21903-5.
- Liu, Fei Tony; Ting, Kai Ming; Zhou, Zhi-Hua (December 2008). "Isolation Forest". 2008 Eighth IEEE International Conference on Data Mining: 413–422. doi:10.1109/ICDM.2008.17. ISBN 978-0-7695-3502-9. S2CID 6505449.
- Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88. JSTOR 2346178.
- Beheshti, N. (2022, March 2) Random Forest Regression. Retrieved October 20, 2022, from https://towardsdatascience.com/random-forest-regression-5f605132d19d