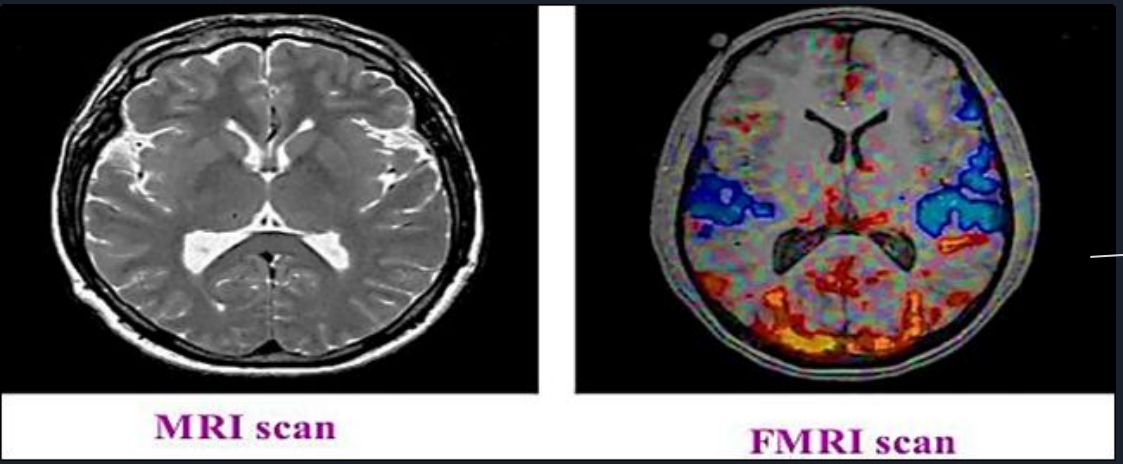
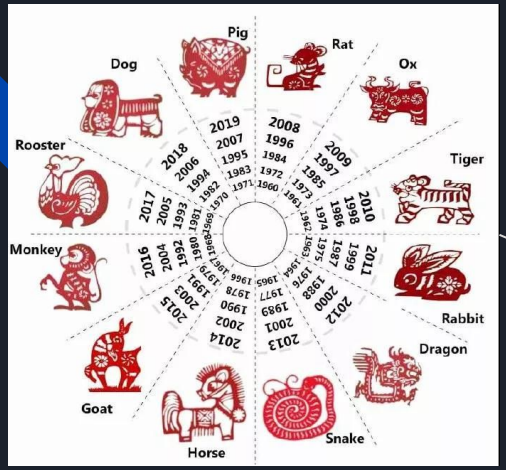
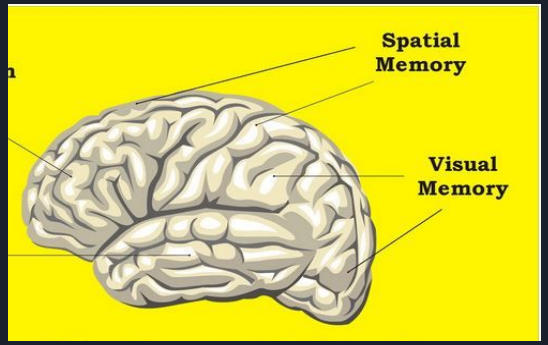
Anderson Group 2
Working memory

Changmin Ahn, Kyungchae Baek, Saul Cervantes,
Maxwell McNeal, Yao Zhang, and Euijun Kim

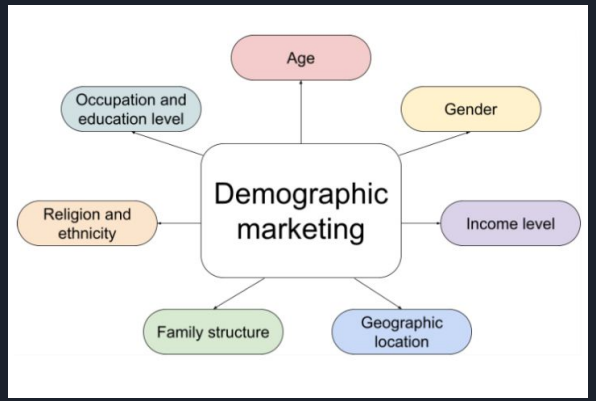


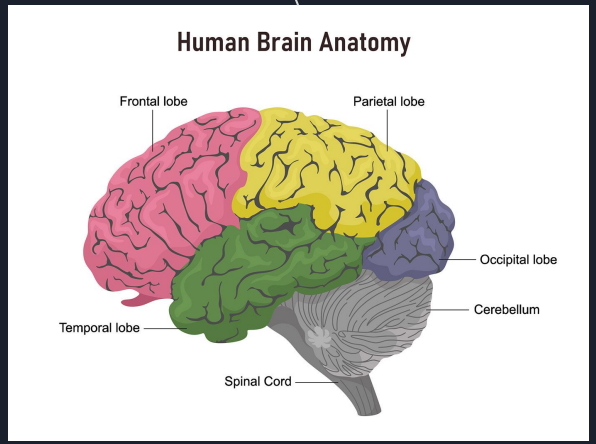
psychiatric disorders





Psychiatric disorders





Other Predictors

Abstract

We can measure brain and brain function using structural and functional MRI. The structural MRI captures a high-resolution 3-dimensional graphic of the brain while the fMRI captures the changes in oxygen levels, an indirect measure of neuronal activity. This type of brain imaging is essential when we investigate the differences among brains with psychiatric disorders. These disorders are highly correlated with demographics and potentially zodiac signs. The study aims to see the best visual and spatial awareness score predictors. We want to know whether demographics or zodiac signs tell us more about brain function than brain imaging data.

Statement of Problem

We would like to find the best visual and spatial working memory score predictors to optimize how we study psychiatry and brain function. Brain imaging data is important, but it is also very expensive and can take lots of time and planning to do. Therefore, with there being large correlations among demographics, zodiacs, and brain function, it would save time and money to determine whether there are variables and predictors that can tell psychiatrists the same or better information that an MRI or fMRI scan can.

Merge and Data Cleaning

- Merge data by PTID
- Removed date columns, duplicate columns, and NAs
- Convert strings and binary columns to factors
- Handling ordinal/nominal variables
- Create Zodiac sign column based on year
- Large data set, contains demographics and response variables, lots of observations
- Small data set, contains MRI and fMRI variables and response variables, more measurements but less observations

Tracking Our Project's Journey

Goal - Examine the extent to which MRI data improve the model prediction

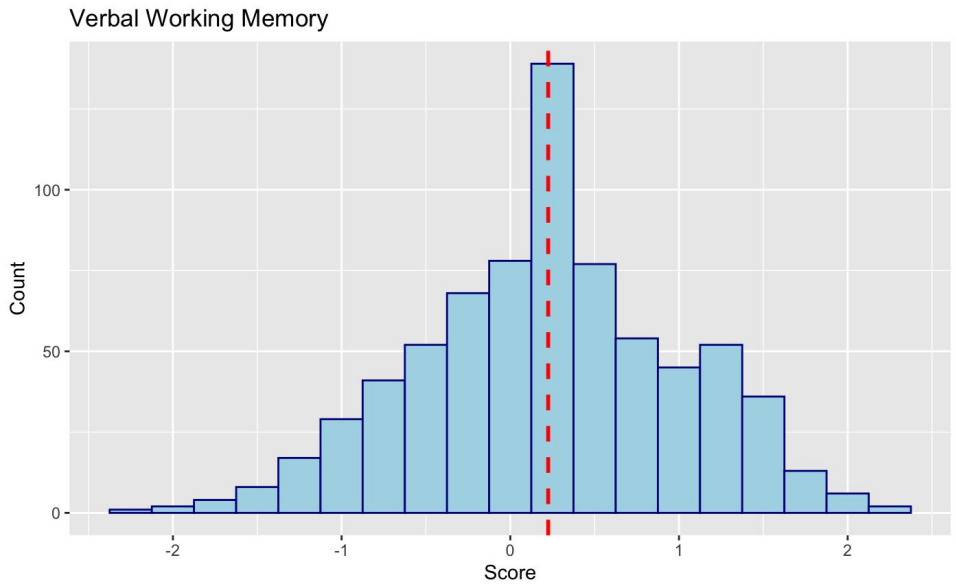
1. Find the most important variables from the all dataset (Large Data) not including MRI data

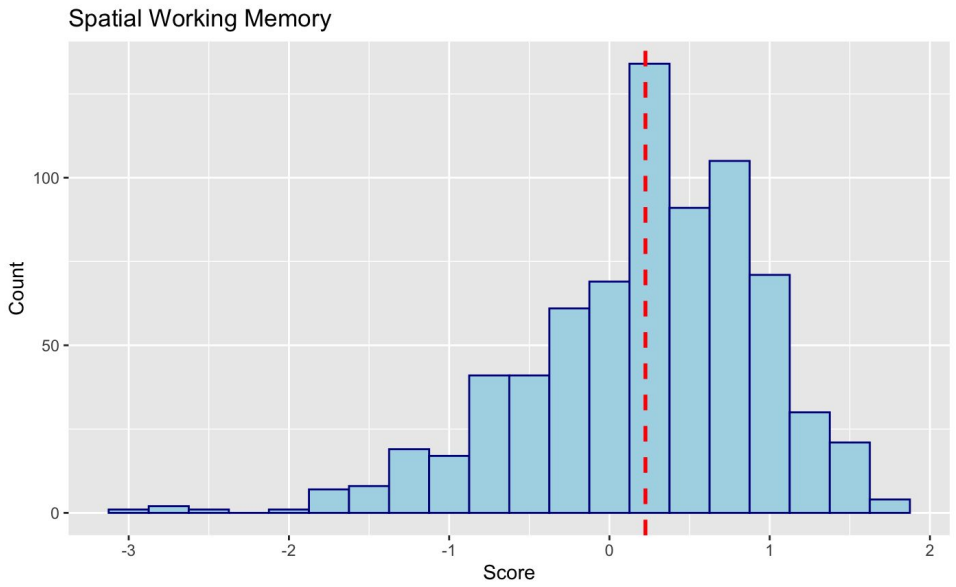
2. Illustrate and test the performance of the model of small data (important variables from Large + MRI)

3. Compare the performances of the large dataset modeling and the small dataset modeling to draw insight from them

EDA Large Data (Without Brain Scans)

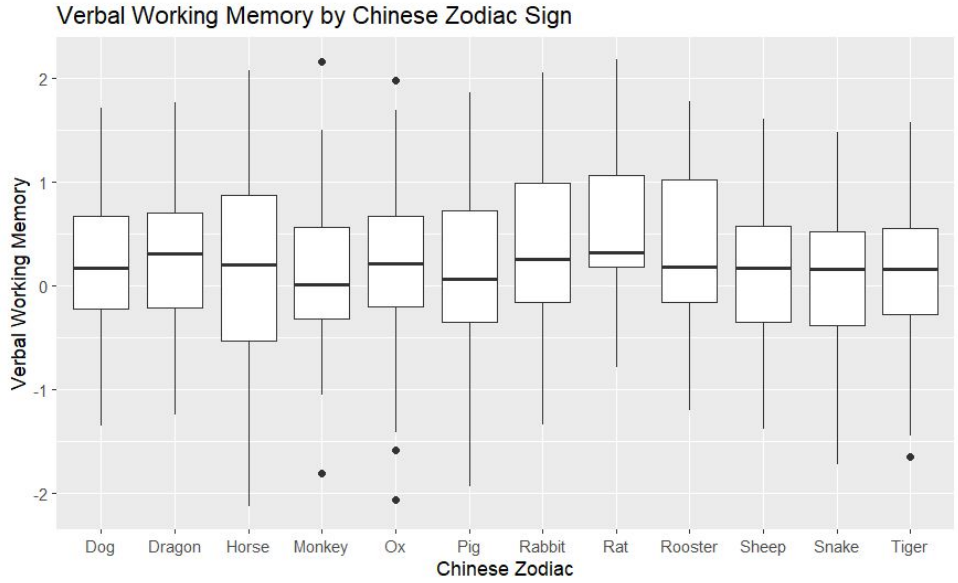
- Most of Verbal working memory values are around 0 and it is very close to normal distribution. Spread from -2 to 2
- Spatial working memory is slightly left-skewed with a similar bulk between -3 and 2

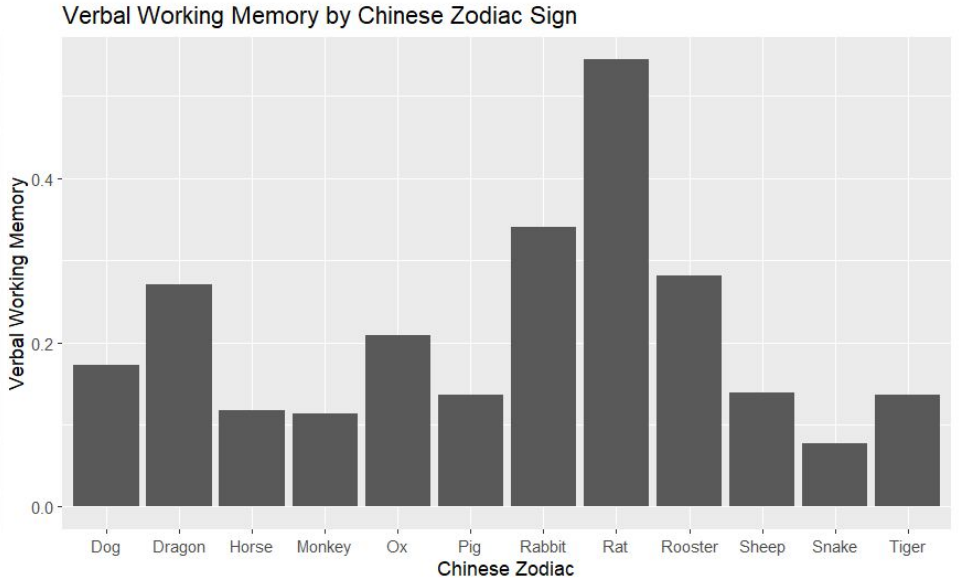




Zodiac Signs and Response Variables - Large data of Verbal Working Memory

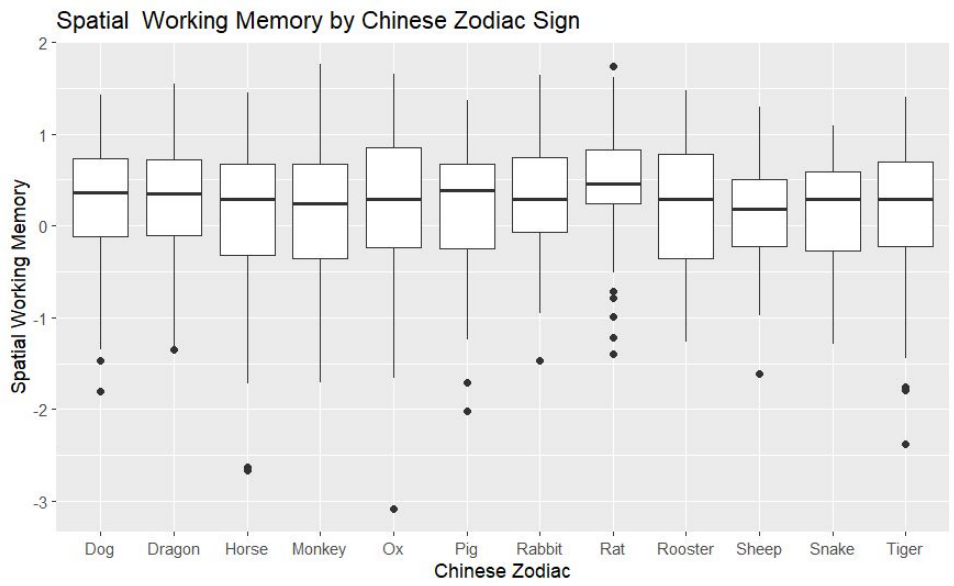
- The mean of each zodiac scores are near to the zero(between 0 to 0.5), and almost all values are land in -2 to 2.

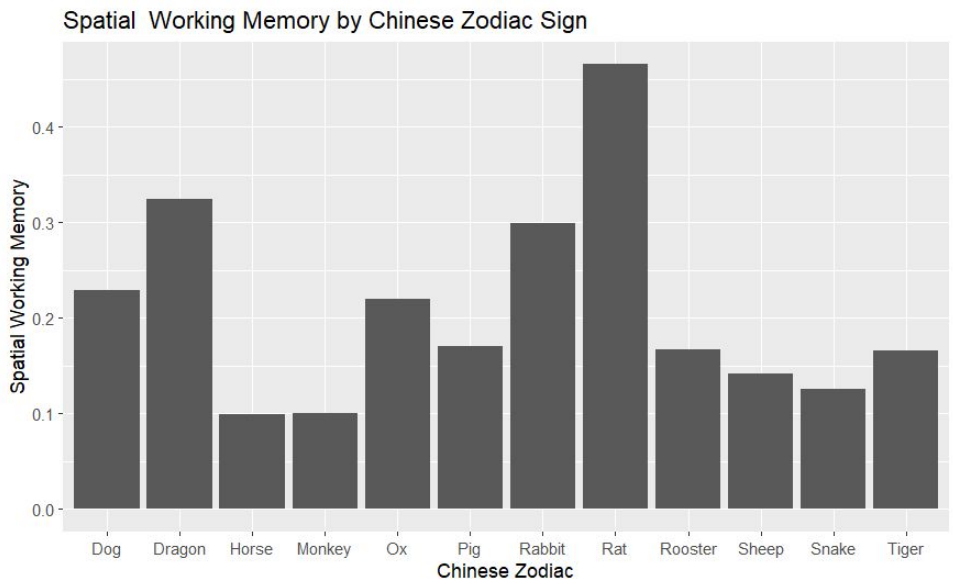




Zodiac Signs and Response Variables - Large data of Spatial Working Memory

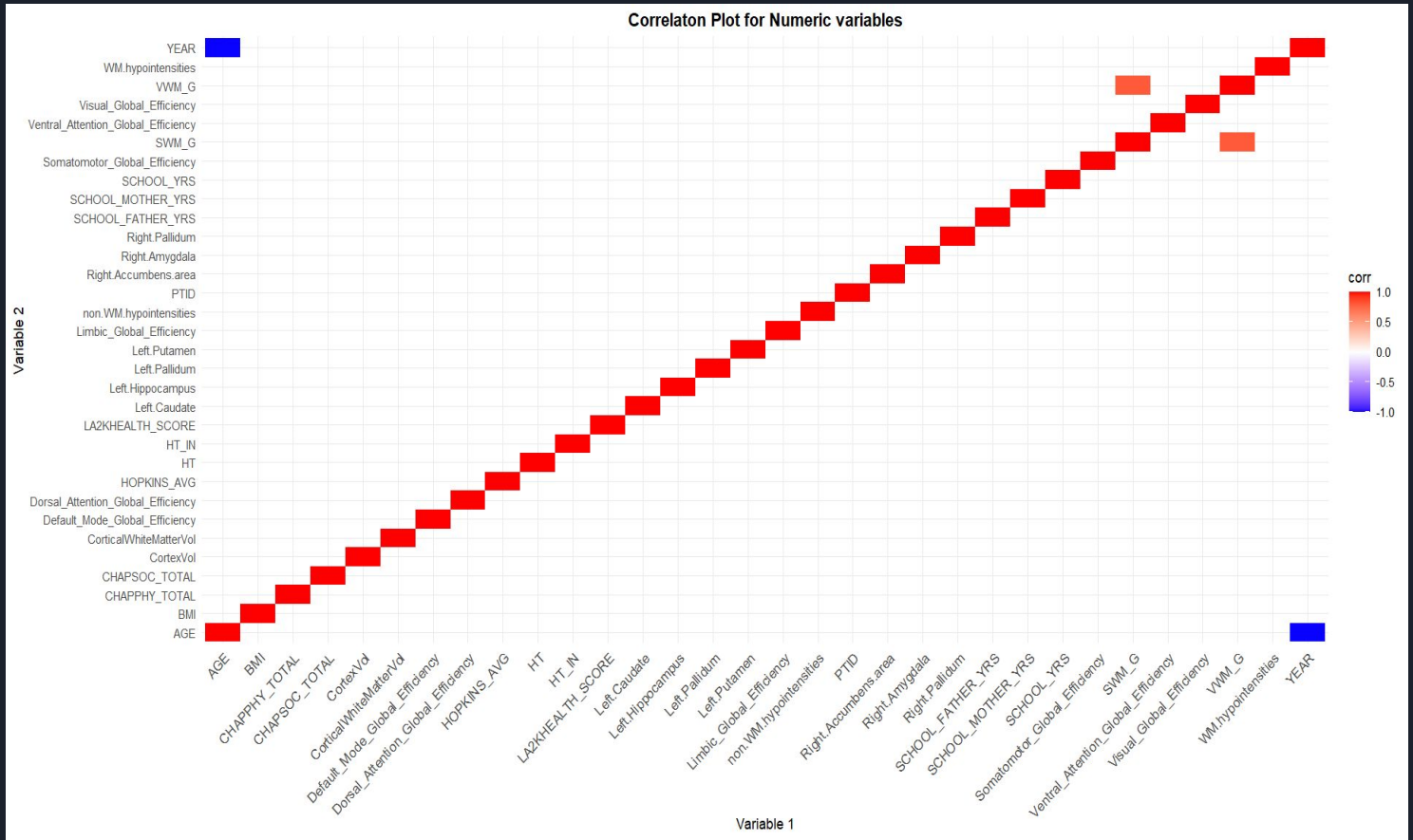
- The mean of each zodiac scores are near to the zero(between 0 to 0.5), and almost all values are land in -2 to 2, but some of points are smaller than -2.





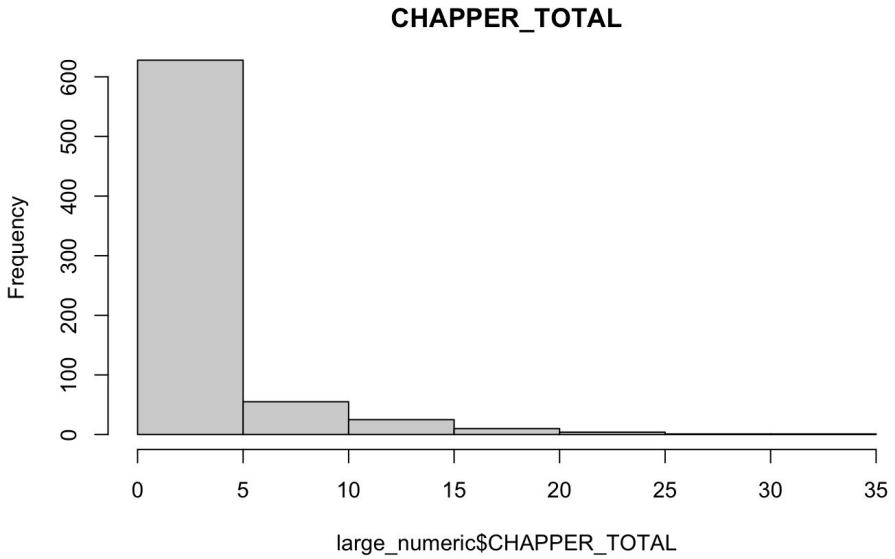
EDA Large Data - Numeric Variables

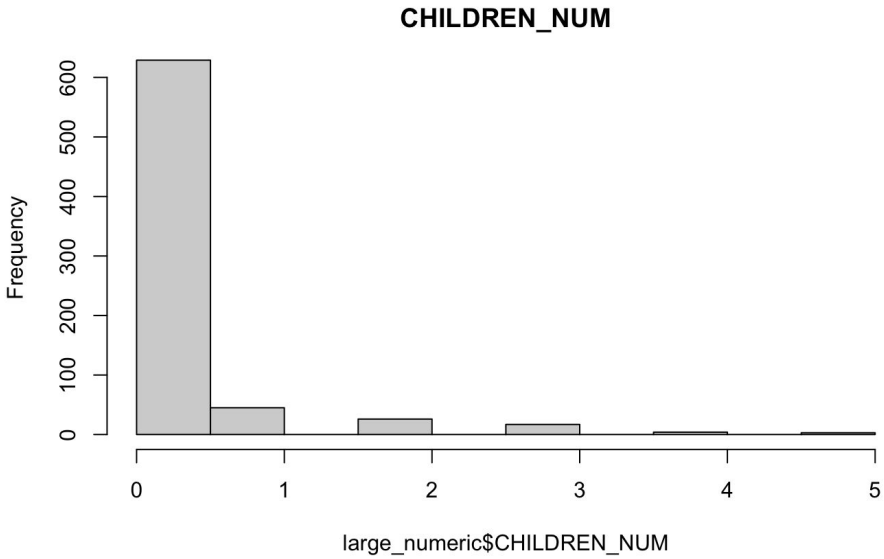
- From the plot, YEAR and AGE are correlated to each other,
- the VWM_G and the SWM_G are moderately correlated to each other.
- No issues with multicollinearity



EDA Large Data - Numeric Variables

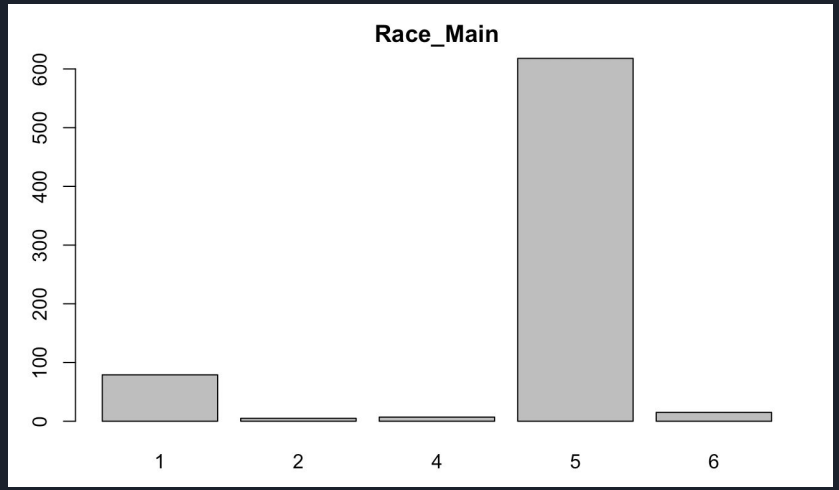
- Nothing unusual about numeric measurements.
- Most are slightly right-skewed or approx normal.
- Highly skewed variables were removed because they would offer little help in a predictive model.





EDA Large Data - Categorical Variables

- Not enough observations in categories to justify using them in a predictive model.
  - Adopt
  - birth-loc
  - Language
  - race_main,
  - Civil_stat
  - Military
  - School_back
  - sexuality
  - cigs



AIC

- We tested AIC with both forwards and backwards stepwise selection.
- Backwards yielded a lower AIC, and thus a better result, than forwards for both spatial and verbal memory.

The resulting model for verbal working memory consists of

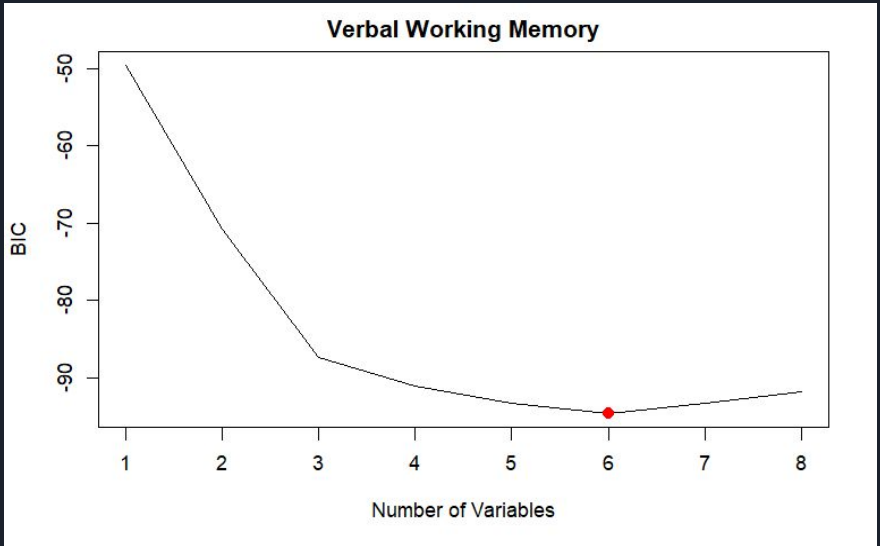
- Age
- Gender
- Ethnicity
- Residence
- School_Years
- School_Mother_Years
- School_Father_Years
- CHAPPHY_Total

The resulting model for spatial working memory consists of:

- Age
- Gender
- Ethnicity
- Residence
- School_Degree
- School_Mother_Years
- CHAPSOC_Total
- CHAPPHY_Total

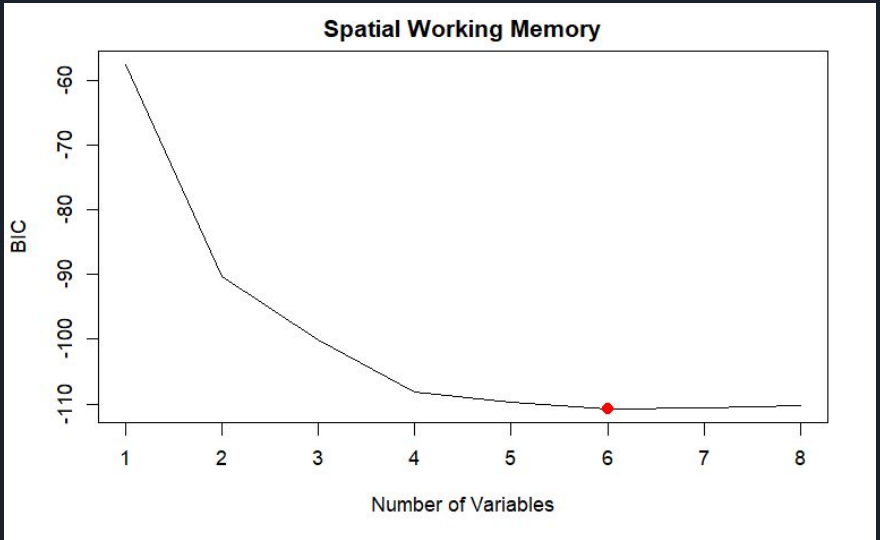
- However, both of these models resulted in high mean squared error, so we believe that we can find better models with other methods.

BIC - Forwards - Large Data



Variables in the model are:

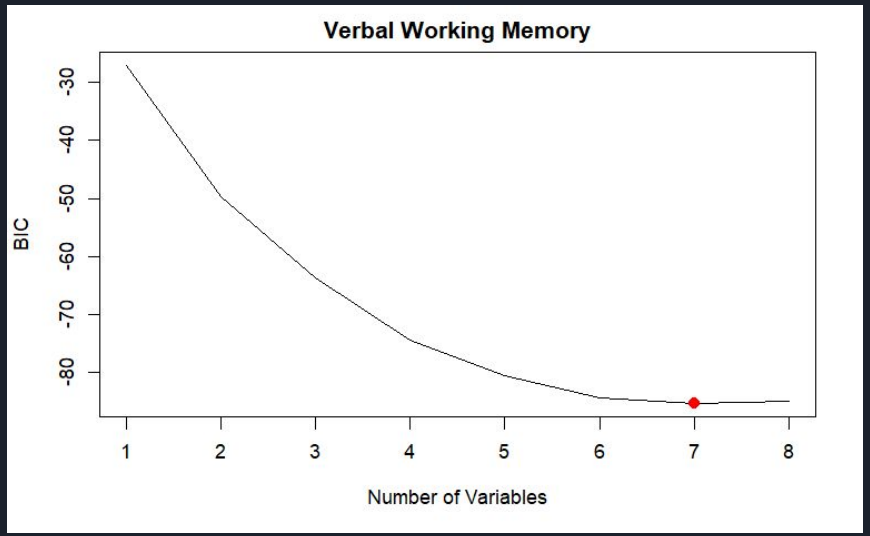
Age, Ethnicity, School_Years, Zodiac, Residence, School_Mother_Degree



Variables in the model are:

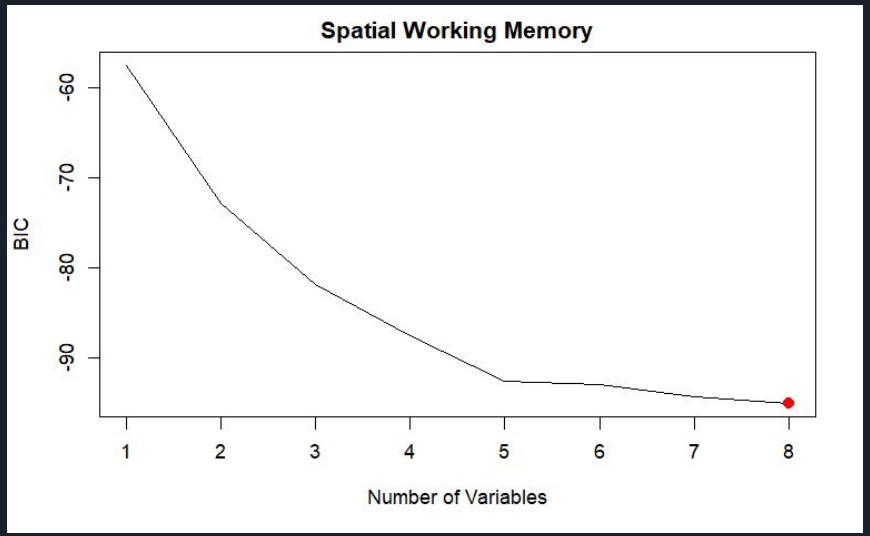
Age, Ethnicity, School_Years, Zodiac, Residence, School_Mother_Degree

BIC - Backwards - Large Data



Variables in the model are:

Age, Zodiac, Father_Ethnicity_Main, Mother_Ethnicity_Main, School_Mother_Degree, School_Degree2, School_Degree5



Variables in the model are:

Age, Gender, Father_Ethnicity_Main, Mother_Ethnicity_Main, School_Degree4, School_Degree5, School_Degree6, School_Degree8

Statistical Modeling/Techniques

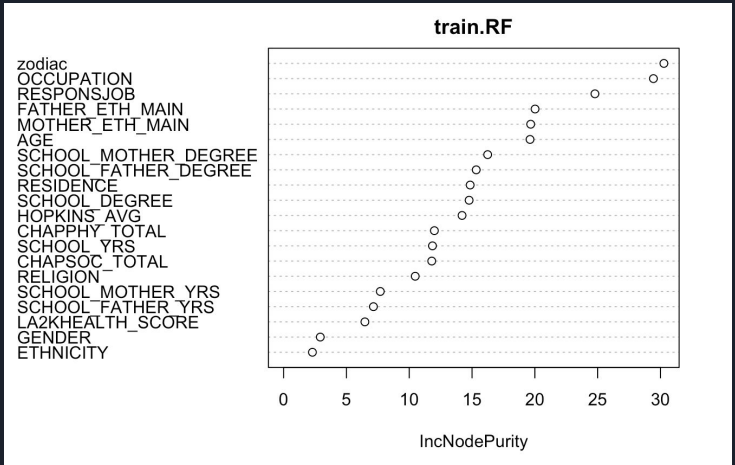
We tried all techniques listed below to perform both variable selection and model performance evaluation:

- Linear Regression model w/ Cross Validation
- XGBoost
- Ridge and Lasso Regression
- K-Nearest Neighbor(KNN)
- Random forest with 100 trees
- Random forest with 500 trees

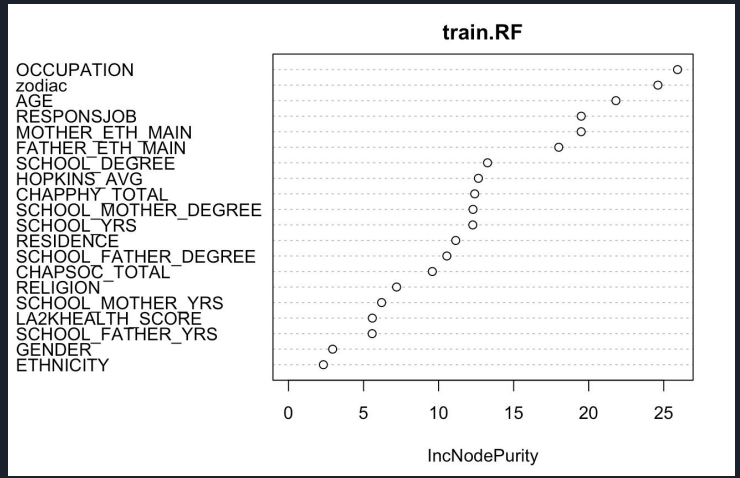
Random Forest with 500 trees - Large Data

Variable Importance Plots

Verbal Working Memory



Spatial Working Memory



Important predictors of both verbal and spatial working memory from the random forest models are:

Age, Occupation, Zodiac, ResponsJob, Mother_Ethnicity_Main, Father_Ethnicity_Main

Best Model for Large Data

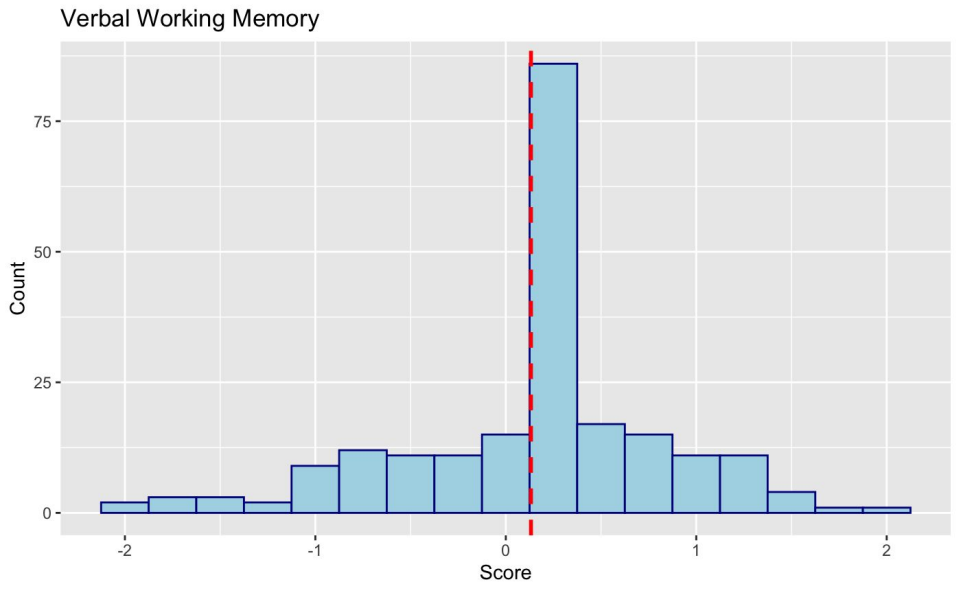
- Comparing the values of RMSE of these methods, we would choose the random forest model with 6 important predictors.

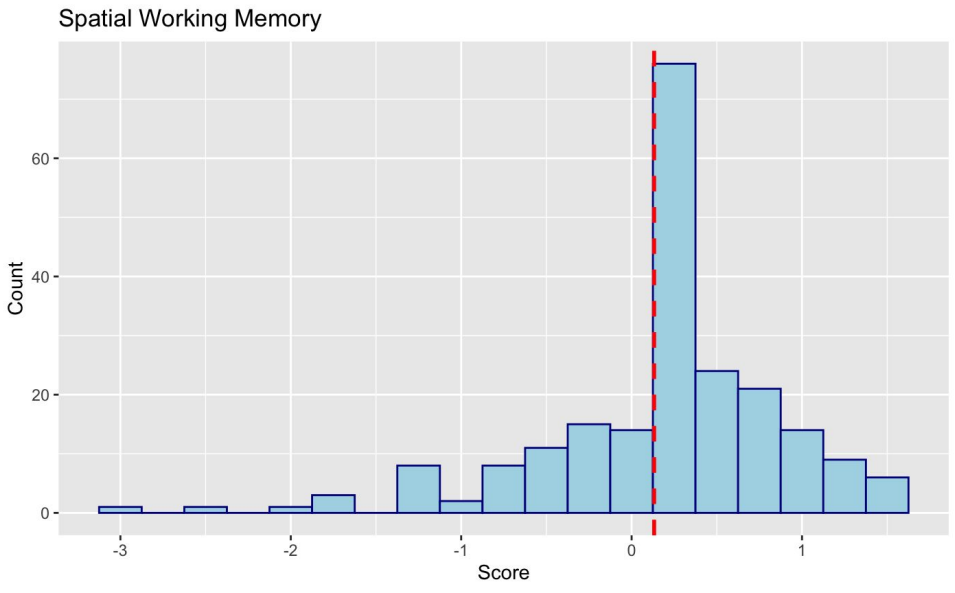
	MLR w/CV	XGBoost	Ridge/Lasso	KNN	RF(100)	RF(500)
RMSE	0.7820632	0.7697241	0.7183195	1.00445	0.7204025	0.693388

- These predictors are
  - Zodiac
  - Occupation
  - ResponsJob
  - Age
  - Father_Ethnicity_Main
  - Mother_Ethnicity_Main

EDA Small Data

- Verbal working memory is symmetric and unimodal with a bulk of its observations between -2 and 2.
- Spatial working memory is slightly left-skewed with a similar bulk between -2 and 2
- Centers are very similar

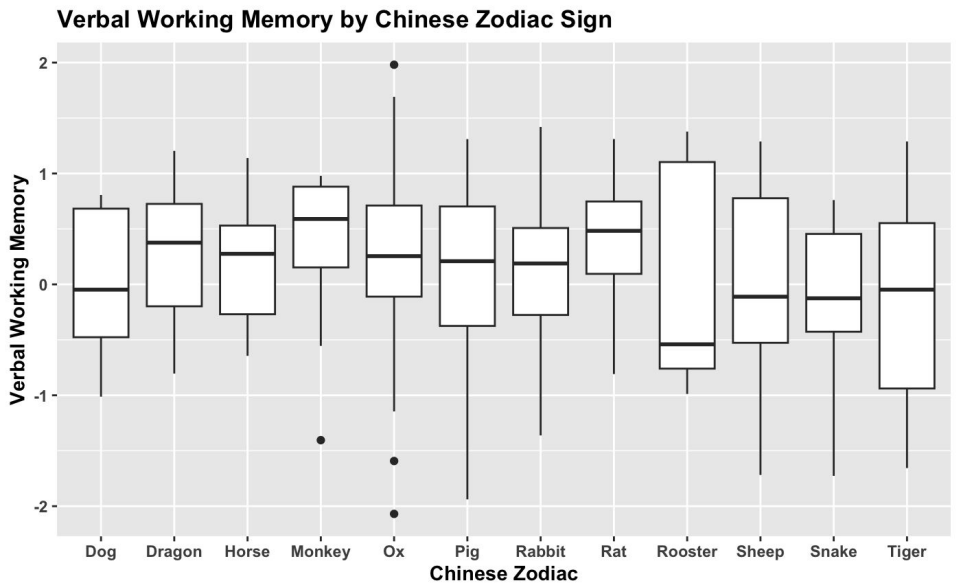


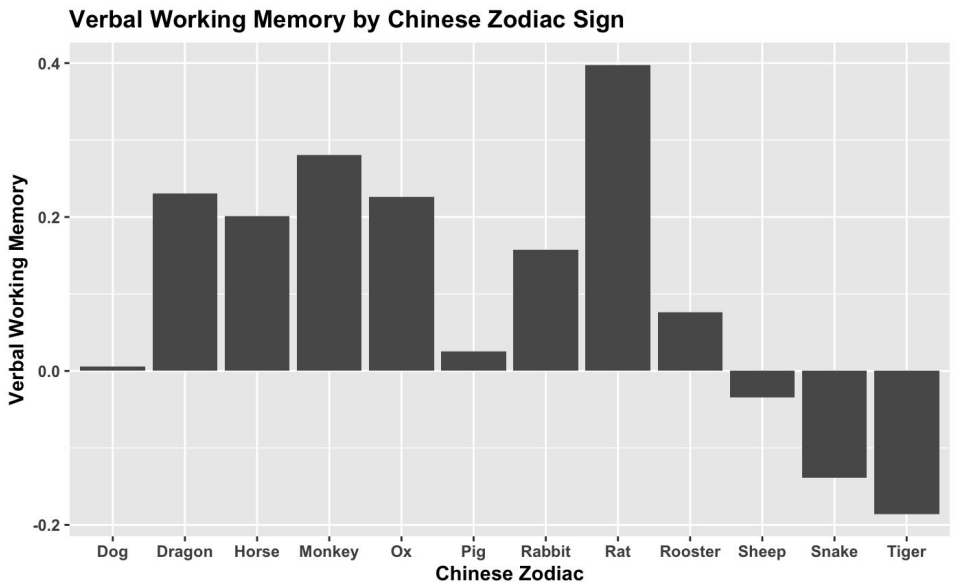


Zodiac Signs and Response Variables

Verbal Working Memory

- Sheep, Snake, and Tiger score less more often

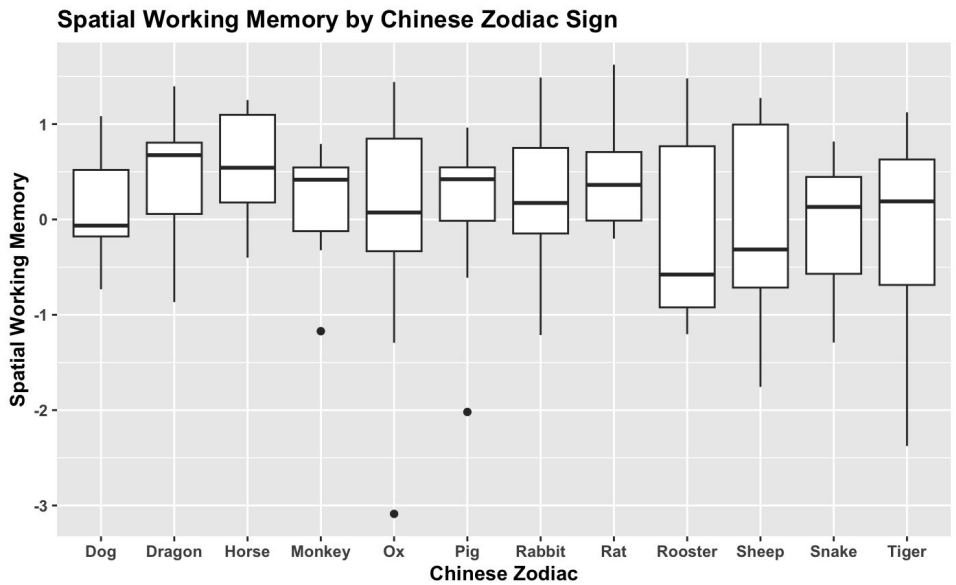


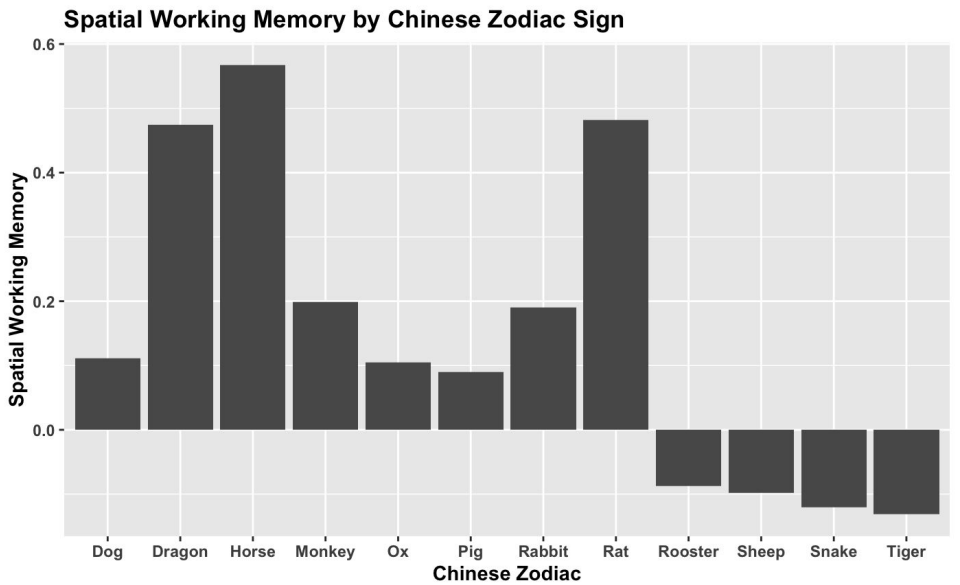


Zodiac Signs and Response Variables

Spatial Working Memory

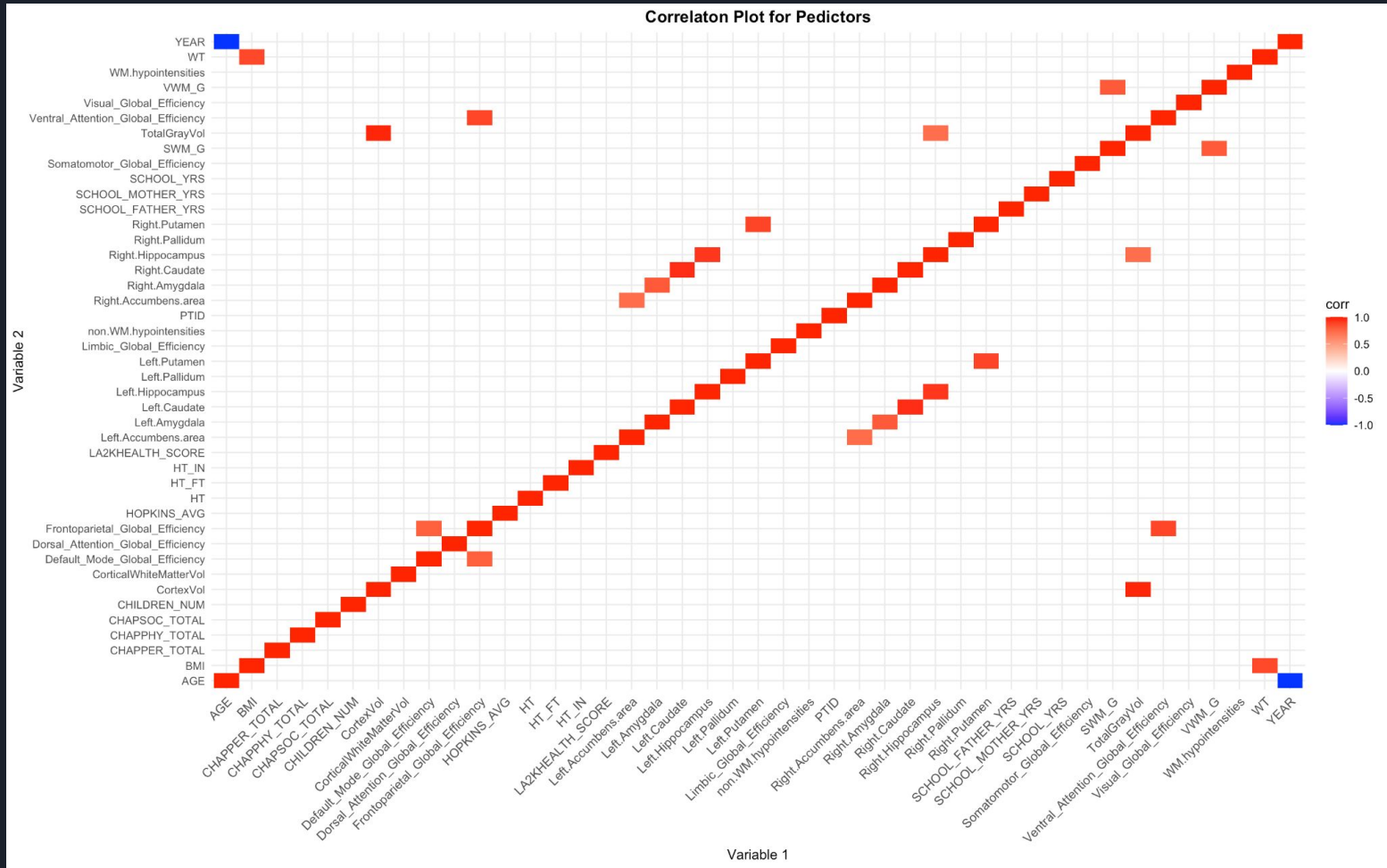
- Rooster, Sheep, Snake, Tiger score less more often





Checking for Multicollinearity - Small Data

- Many variables are positively correlated to each other.



Numeric and Categorical Variables - Small Data

- There were no extremely skewed or unbalanced variables that needed to be removed
- Many correlated variables
  - Most correlated to response were kept while others removed

BMI	WT	Right.Amygdala	Left.Amygdala
Cortex_Vol	TotalGrayVol	Left.Caudate	Right.Caudate
Default_Mode_Global_Efficiency	Frontoparietal_Global_efiiciency	Left.Hippocampus	Right.Hippocampus
Ventral_Attention_Global_efficiency	Frontoparietal_Global_efiiciency	Left.Putamen	Right.Putamen
Right.Accumbens.area	Left.Accumbens.area		

Random Forest - Small Data

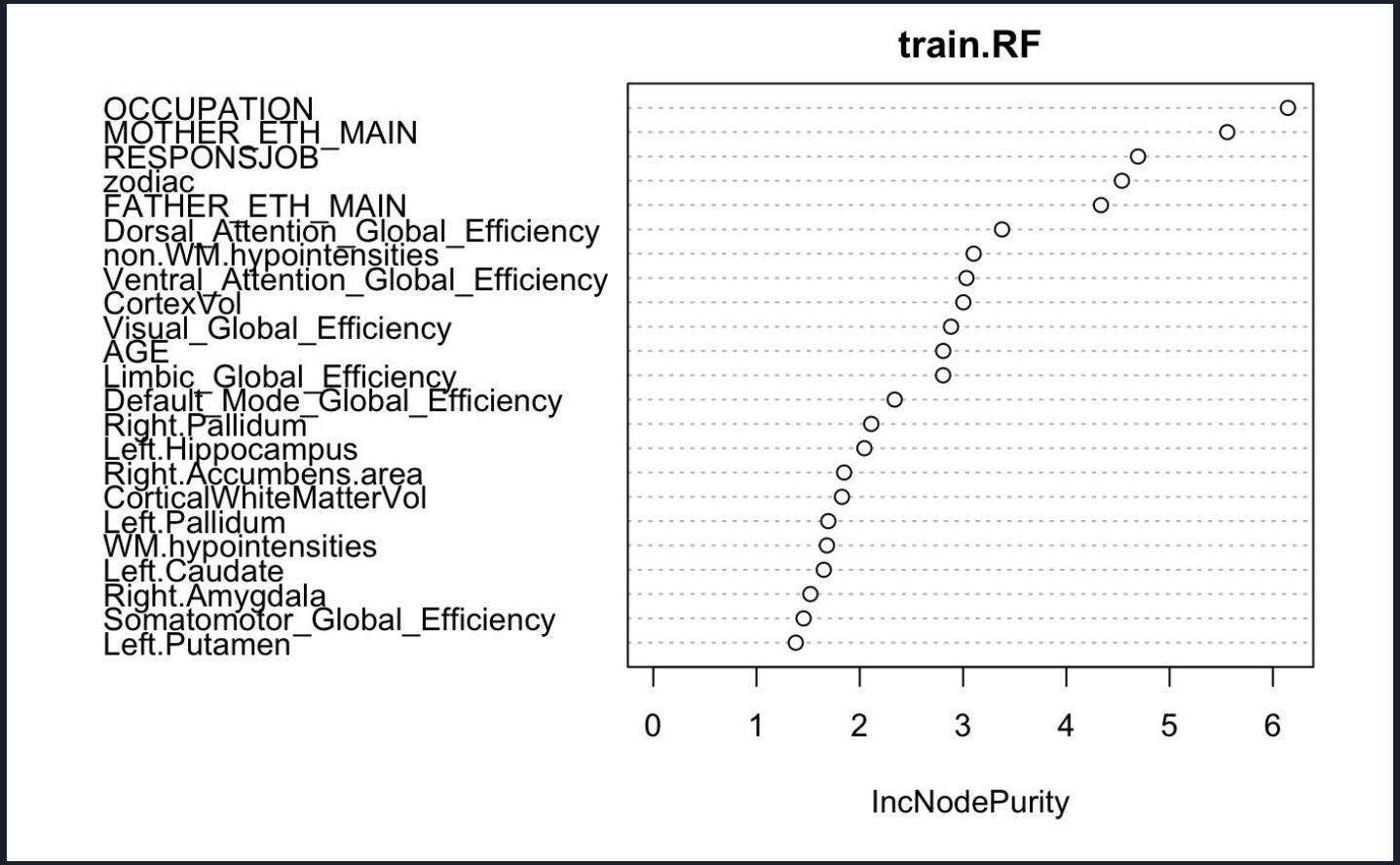
- 100 trees
  - For VWM the best predictors are OCCUPATION, RESPONSJOB, and Zodiac
    - RMSE : 0.75305066
    - Majority of important variables are not MRI measurements
  - For SWM the best predictors are FATHER_ETH_MAIN, MOTHER_ETH_MAIN, OCCUPATION, RESPONSJOB, and Zodiac
    - RMSE : 0.7239158
    - Majority of important variables are not MRI measurements

Random Forest - Small Data

- 500 Trees
  - For VWM the best predictors are OCCUPATION, MOTHER_ETH_MAIN, RESPONSJOB, Zodiac, and FATHER_ETH_MAIN
    - RMSE : 0.6935013
    - Majority of important variables are not MRI measurements
  - For SWM the best predictors are Zodiac, Limbic Global Effieciency, Default Mode Global Efficiency, Age, and MOTHER_ETH_MAIN
    - RMSE : 0.6447348
    - Majority of important variables are not MRI measurements

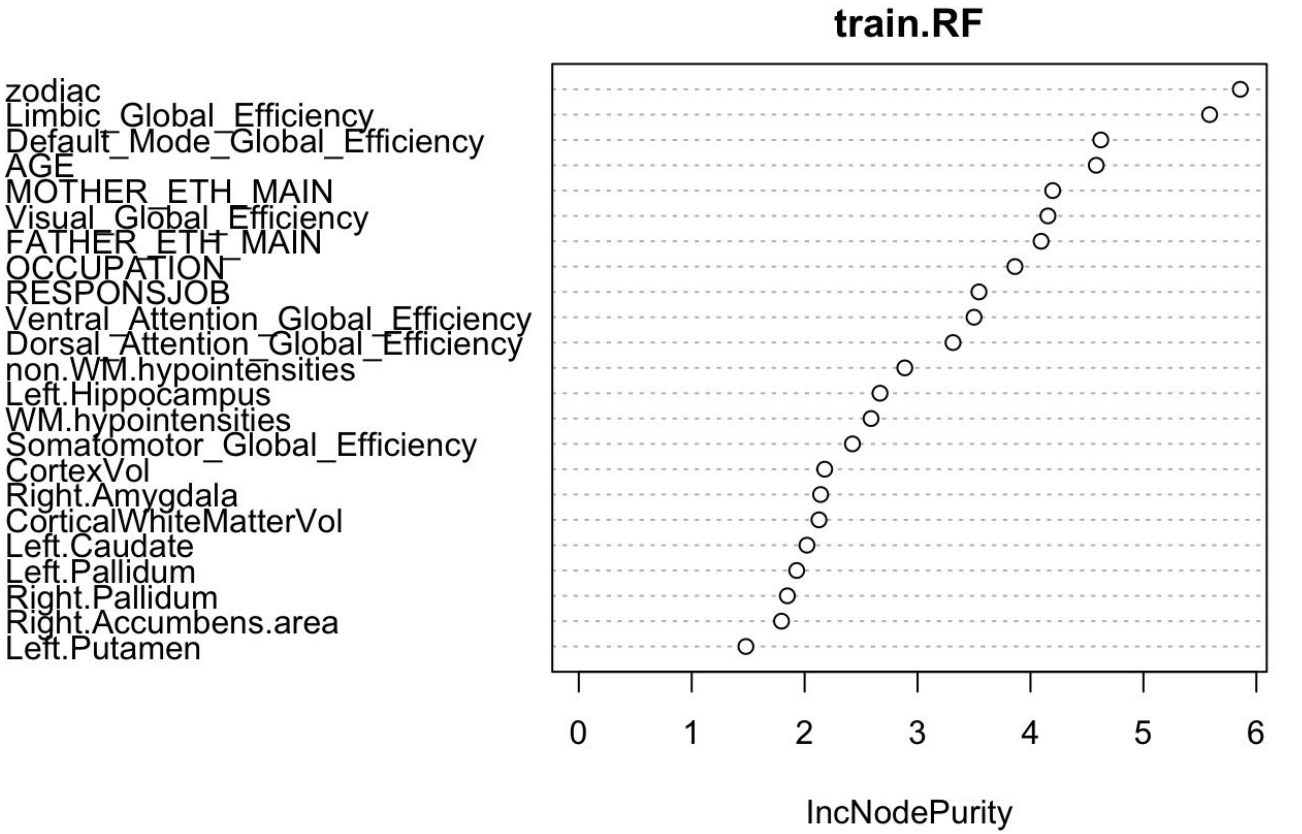
Random Forest - Small Data

- 500 trees for Verbal Working Memory



Random Forest - Small Data

- 500 trees for Spatial Working Memory



Only With MRI variables

	MLR	KNN	RF (500 trees)
RMSE	0.74	0.73	0.69

- Testing a random forest model with only MRI variable doesn't give better results.
- The R-squared value is low with 0.18, enough to not say it is a good predictor.

Thoughts After Variable Selection

- MRI variables may not be as important for predicting memory
- The non-MRI variables are already doing a job of predicting memory, and the MRI variables are providing some additional predictive power, but not enough to make a significant difference.

Conclusion

- The demographic variables in the large dataset are strong predictors of both verbal and spatial working memory, such that the addition of MRI variables in the small dataset does not substantially improve the model's performance.

- This leads us to believe that the demographic variables are capturing most of the relevant information about memory, thus the MRI variables are not adding substantially more prediction power about memory.

- This does not mean that the MRI variables in the small dataset do not provide valuable information about verbal and spatial memory. It instead indicates the demographic variables are strong enough predictors on their own that the addition of MRI variables does not substantially improve the model's performance.

Shortcomings - Things To Improve

- The demographic data set has many more observations than the brain imaging set (714 rows versus 214 rows). This imbalance in sample size affects our final models.
- If we had more observations in the brain imaging data set, our model using the brain imaging data could improve and possibly perform significantly better than our final model using the demographic data.
    - However, gathering more data would be expensive and time consuming.
- We should investigate if zodiac is simply masking the importance of birth year. Is there a significant event that could affect people's working memory more or less depending on their birth year?
    - For example, the COVID-19 pandemic

Thank you!!!

Answer to Dr Anderson's question

Based on our analysis, there appear to be differences in working memory among factor variables such as ethnicity, but due to limited observations, it is difficult to trust that these differences are truly significant, which is why we ended up opting not to divide our small dataset into two different groups (patients and controls). Instead, we combined the patient and control groups to increase the number of observations and improve our ability to predict working memory.

Although there is a strong intuition that there may be differences in working memory between males and females, the limited amount of data available prevents us from making any conclusive statements regarding this potential difference. However, it is possible that further analysis could reveal differences in specific predictors.

Additionally, the small data for the patient and control groups performed similarly or slightly better in predicting working memory compared to the combined dataset that was not divided into groups. However, as we have not enough observations to trust the outcome, we rather chose to conduct analysis without dividing the dataset, focusing on whether MRI data helps prediction or not.