USING VISTA DATA TO ASSIGN TRIPS TO SYNTHETIC POPULATION

THE VICTORIA INTEGRATED SURVEY OF TRAVEL AND ACTIVITY(VISTA) SURVEY GIVES TRIPS AND ACTIVITIES INFORMATION ABOUT 1% OF THE POPULATION OF VICTORIA. THIS INFORMATION INCLUDES THE DETAILS OF THE PERSON (AGE, SEX, PROFESSION ETC.) AND THE TRIPS HE/SHE TOOK ON A GIVEN DAY. THE DETAILS OF THE ORIGIN AND DESTINATION OF THESE TRIPS, THE MODE OF TRAVEL, PURPOSE OF TRIPS ETC. ARE KNOWN. USING CLASSIFICATION AND REGRESSION TECHNIQUES, WE AIM TO PREDICT THE KIND OF TRIPS TAKEN BY A LARGE SYNTHETIC POPULATION, WHICH IS IDEALLY EVERY PERSON LIVING IN MELBOURNE, BASED ON OUR KNOWLEDGE OF THE SMALLER VISTA SAMPLE. IN THIS TALK, THIS WORK WILL DISCUSS THE INITIAL RESULTS FROM THIS EFFORT FOR A SUBURBS IN MELBOURNE

Datasets

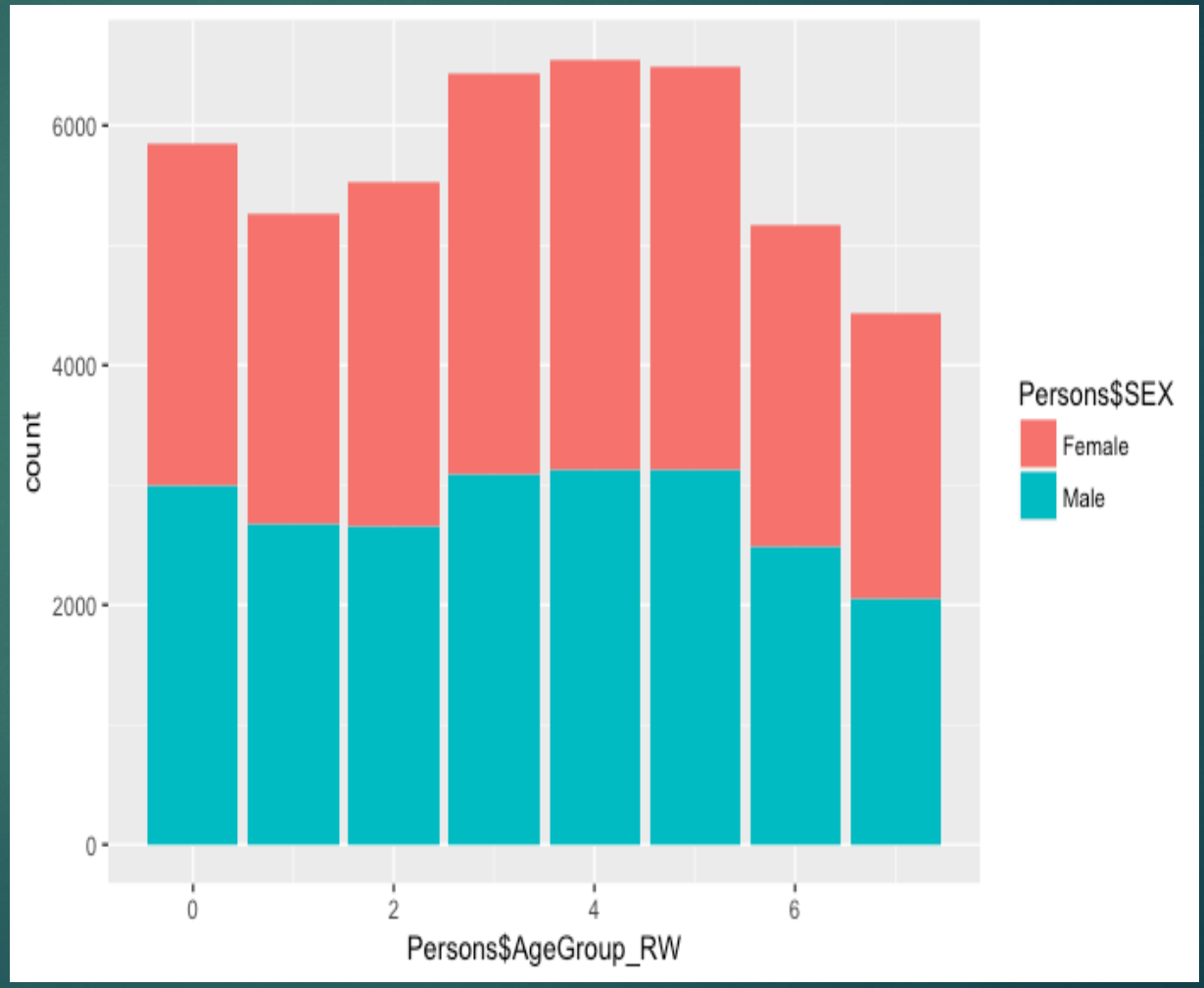
- VISTA DATA

- Persons
- Households
- Trips
- Stops
- Journey to Work
- Journey to Education

- Synthetic Population
  - Divided across 308 suburbs.

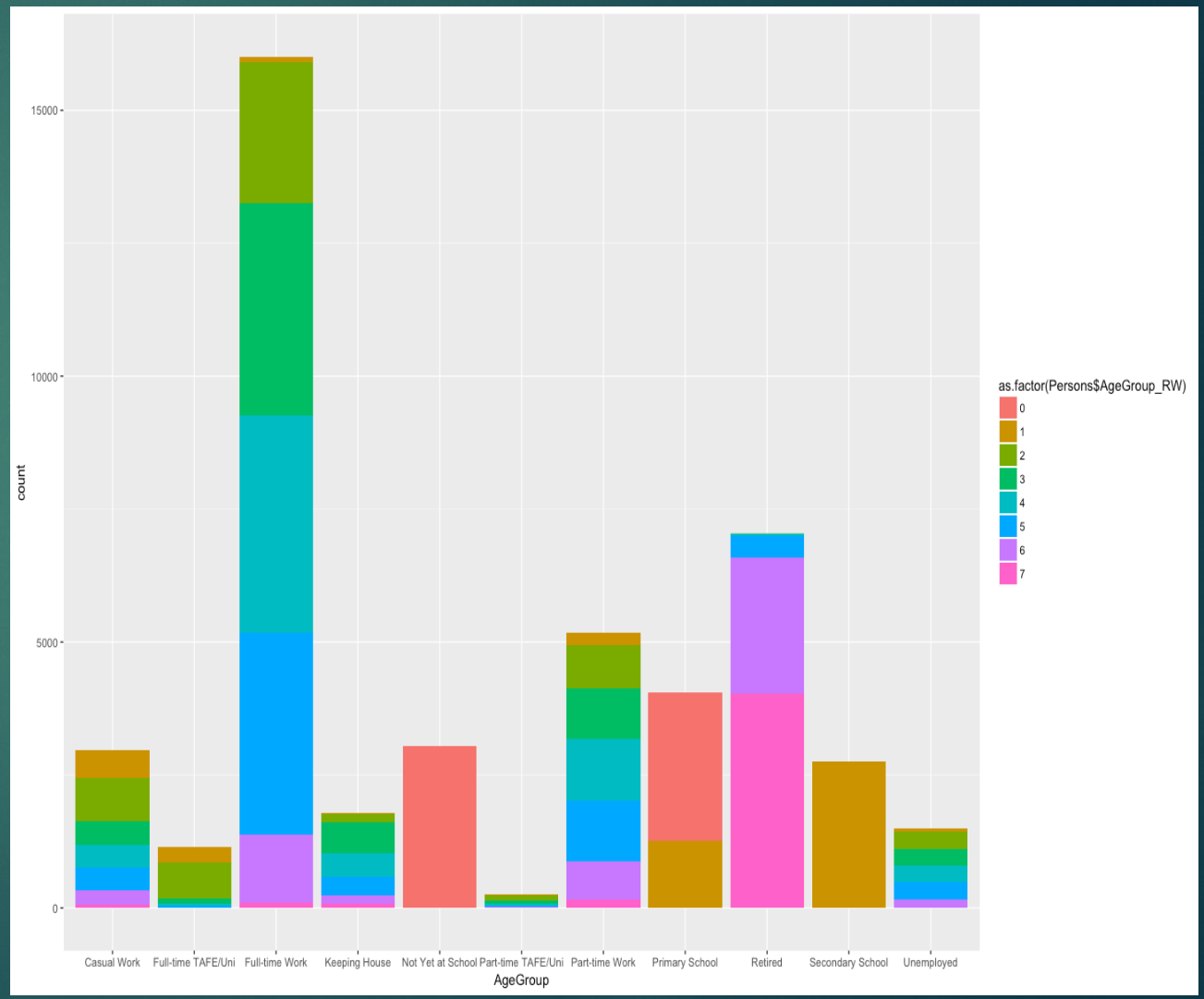
Important Variables

- PersID
- Age
- AgeGroup
- Sex
- MainAct
- ANSZCO1
- Weights
- ModeOfTravel
- Distance
- Time
- Speed



Persons – MAIN ACT

	MainAct	Frequency
1	Casual Work	2961
2	Full-time TAFE/Uni	1143
3	Full-time Work	16010
4	Keeping House	1779
5	Not Yet at School	3053
6	Part-time TAFE/Uni	252
7	Part-time Work	5171
8	Primary School	4045
9	Retired	7042
10	Secondary School	2759
11	Unemployed	1499



Persons- ANSZCO1

Very Important for Work force . Different professions have different start times, durations etc.
There is a huge chunk of population in the 'Not in Work Force' section. Later, we have assigned them one activity which they are **Most Likely** to be doing as their ANSZCO1.

	0	1	2	3	4	5	6	7
Clerical and Administrative Workers	0	45	441	483	631	674	346	37
Community and Personal Service Workers	0	214	488	394	438	412	180	34
Labourers	0	121	446	431	468	527	236	49
Machinery Operators and Drivers	0	14	101	168	209	263	156	11
Managers	1	18	278	694	845	702	254	40
Missing/Refused	0	5	17	21	23	24	3	1
Not in Work Force	5839	4351	1264	1047	933	1231	2876	4126
Professionals	0	77	1344	2289	2053	1799	737	85
Sales Workers	0	323	626	367	396	365	179	30
Technicians and Trades Workers	0	89	523	539	556	496	210	21

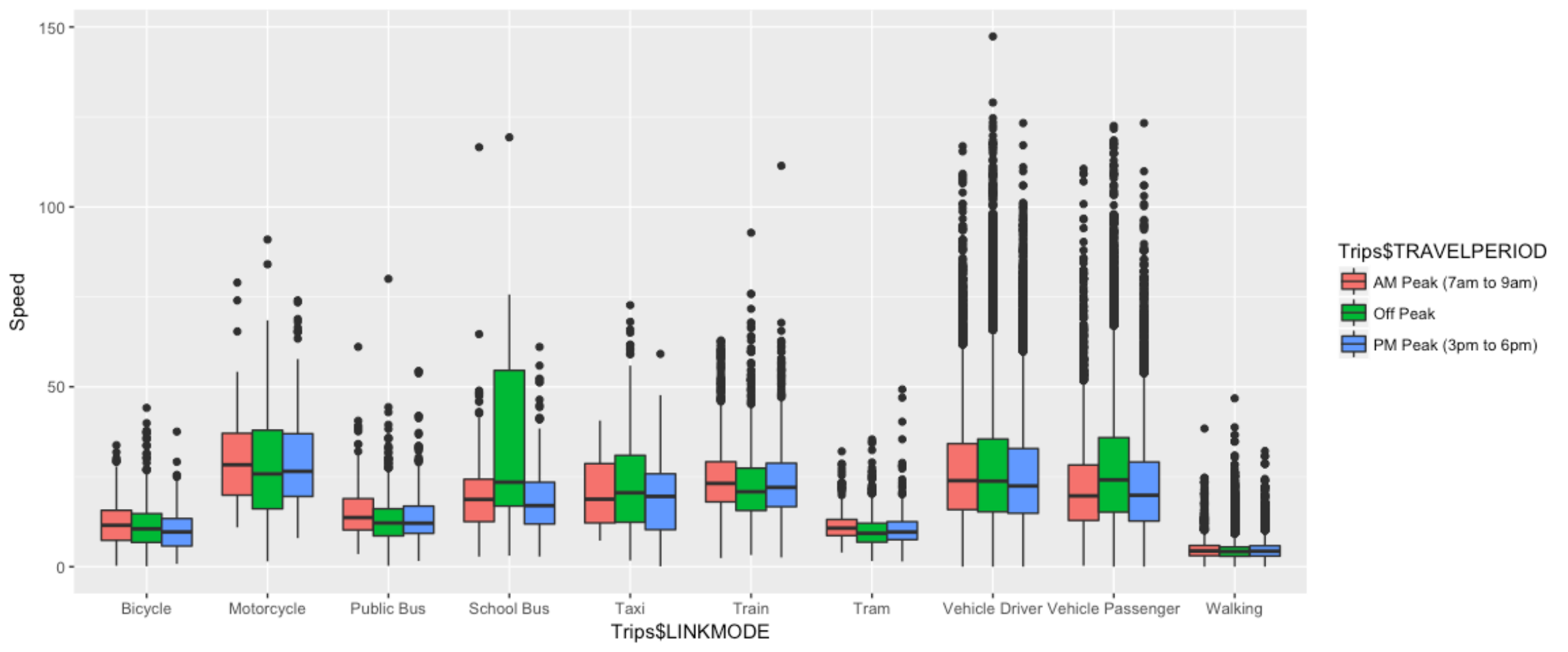
Weights and their meaning

- Weights refer to how many people in Melbourne does one person in the VISTA data represent
- The following is the same table as that of the earlier page, except using weights.
- If we add all of these, we will get the population of Melbourne.

	0	1	2	3	4	5	6	7
Clerical and Administrative Workers	0.00	4632.74	58232.47	52530.30	59806.99	54830.79	26942.24	3410.27
Community and Personal Service Workers	0.00	21814.51	65072.60	42308.55	43411.12	34529.67	15181.95	3468.01
Labourers	0.00	12701.56	60594.98	47176.00	47516.29	46146.72	20500.28	5469.21
Machinery Operators and Drivers	0.00	1555.72	14228.70	19545.86	21941.79	23814.98	13079.99	1099.23
Managers	93.97	1701.79	37401.39	76941.23	78715.42	56230.62	21362.25	4115.85
Missing/Refused	0.00	385.53	2166.11	2333.48	1974.65	2058.20	241.77	72.38
Not in Work Force	586884.92	447613.98	173582.41	113011.87	92937.16	106569.95	235224.74	407302.35
Professionals	0.00	7677.60	179311.73	253371.78	192633.26	145543.59	59293.76	8742.56
Sales Workers	0.00	31985.06	80778.42	40053.77	37731.50	30528.62	13989.66	2684.96
Technicians and Trades Workers	0.00	9286.02	68938.54	57883.27	54489.14	41646.88	17331.04	2072.42

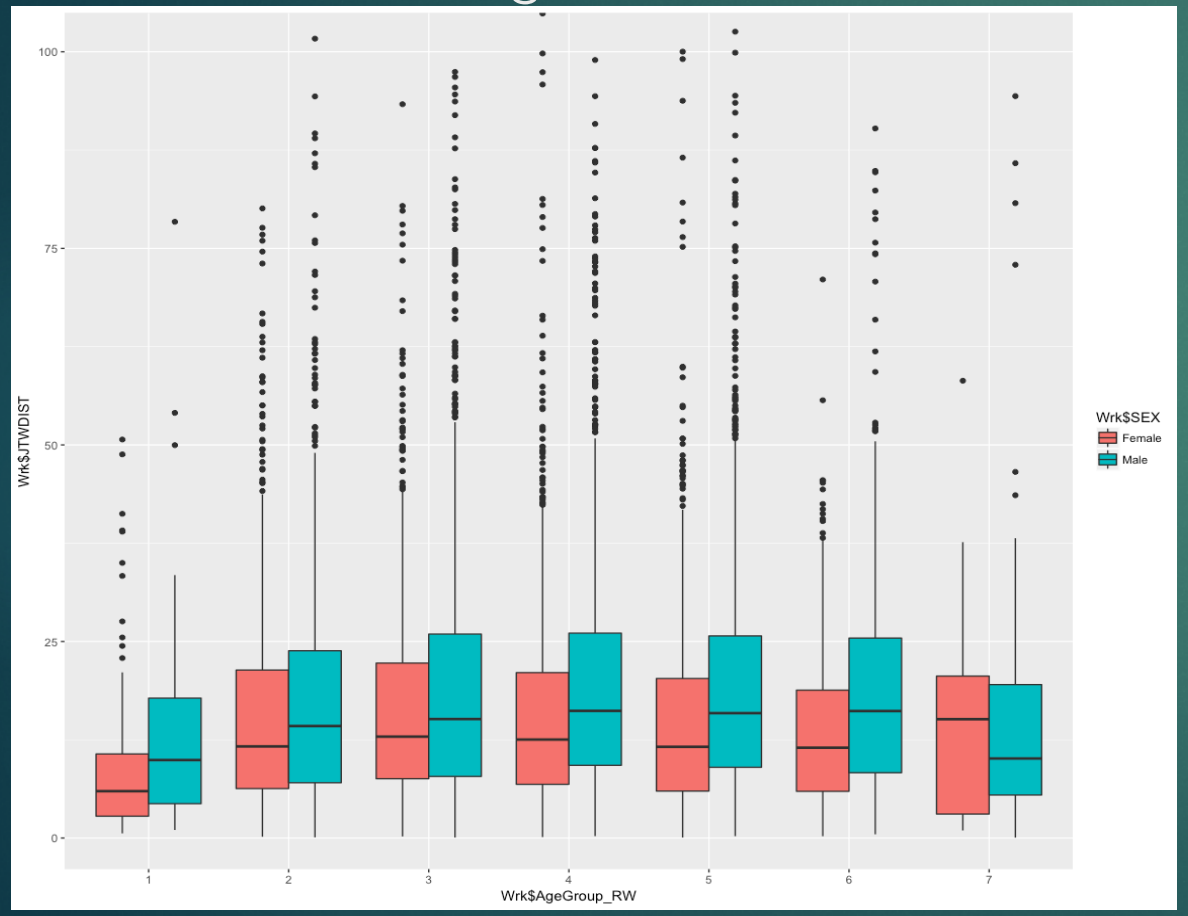
Trips – LinkMode

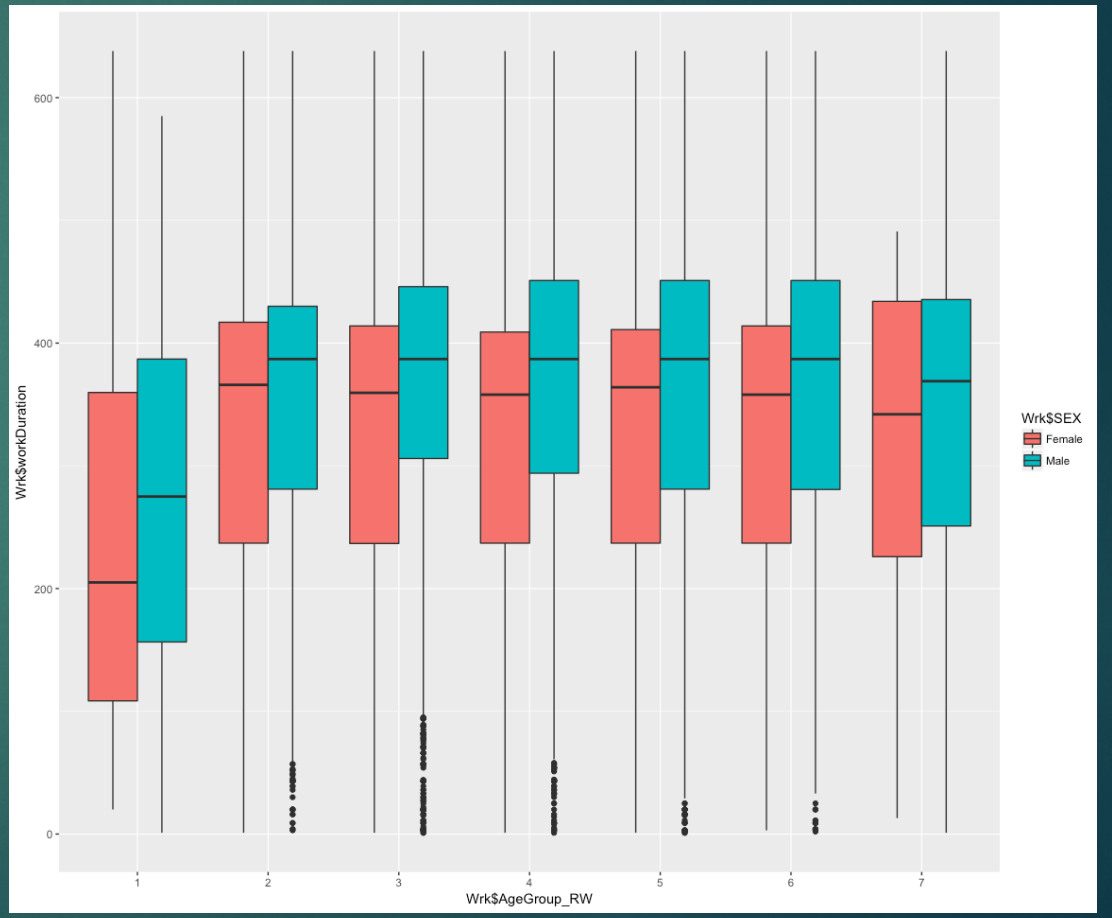
While many modes may be used in a trip, LinkMode is the Major Mode of Travel



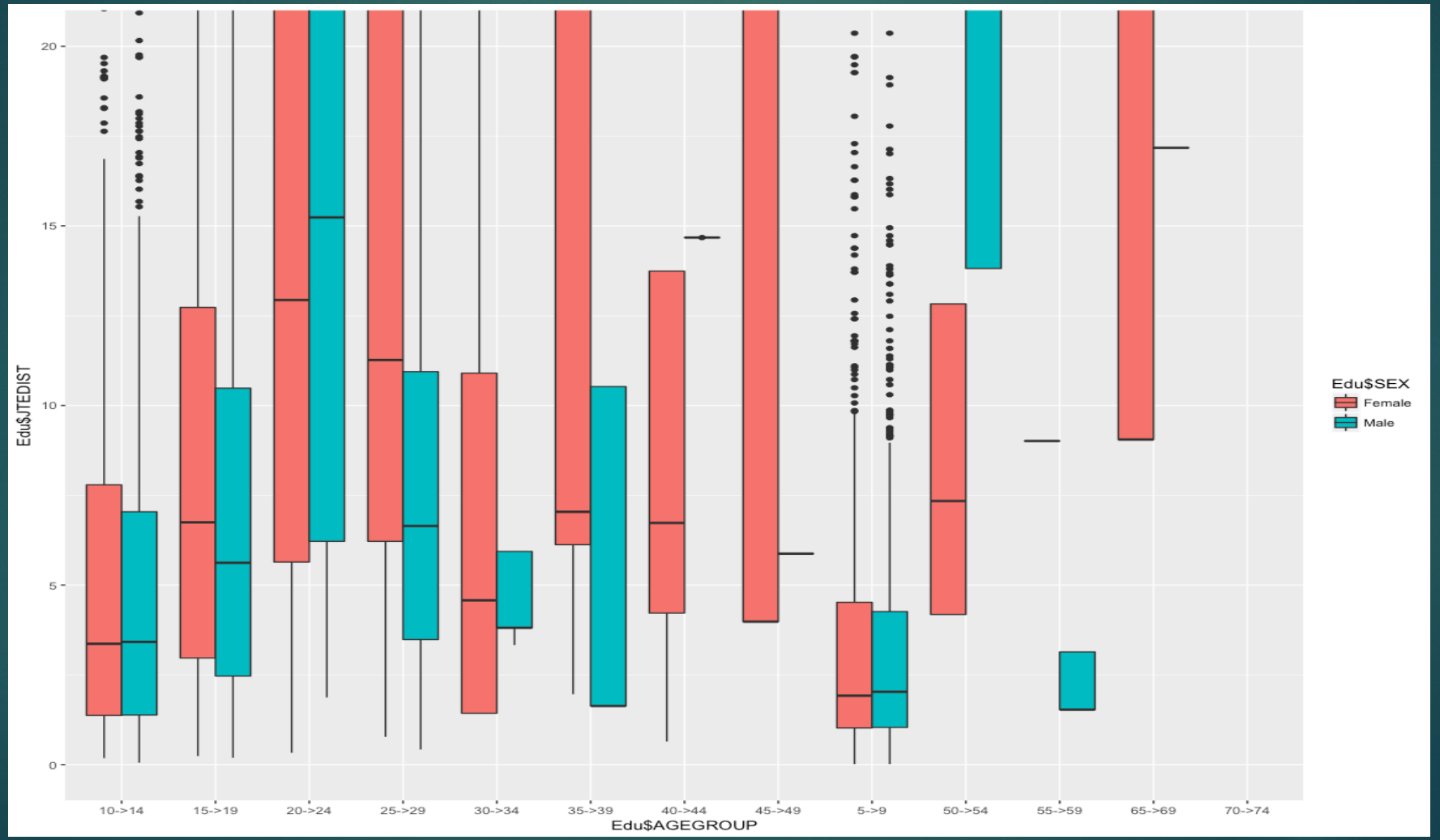
Journey To Work

We have merged the Trips dataset with Persons dataset to get all the information together





Journey To Education



SYNTHETIC POPULATION

In all these examples, we will be examining the dataset of Northcote

The same code can be used for each suburb (explained later)

Current Synthetic Data

AgentId	Age	Gender	RelationshipStatus	PartnerId	MotherId	FatherId	ChildrenIds	RelativeIds
206021111P42266	24	Male	MARRIED	206021111P47217			[206021111P56065, 206021111P64793, 206021111...	
206021111P42267	23	Male	MARRIED	206021111P47218			[206021111P57292, 206021111P64807]	[206021111P52277]
206021111P42268	21	Male	MARRIED	206021111P47219			[206021111P56928]	
206021111P42269	22	Male	MARRIED	206021111P47220				
206021111P42270	16	Male	MARRIED	206021111P47221			[206021111P55573, 206021111P64988]	
206021111P42271	21	Male	MARRIED	206021111P47222			[206021111P53828]	
206021111P42272	24	Male	MARRIED	206021111P47223				
206021111P42273	24	Male	MARRIED	206021111P47224			[206021111P56949]	
206021111P42274	18	Male	MARRIED	206021111P47225				
206021111P42275	18	Male	MARRIED	206021111P47226				
206021111P42276	24	Male	MARRIED	206021111P47227			[206021111P56147]	
206021111P42277	18	Male	MARRIED	206021111P47228			[206021111P55870, 206021111P64714]	
206021111P42278	22	Male	MARRIED	206021111P47229			[206021111P55153, 206021111P64859]	[206021111P64684]
206021111P42279	16	Male	MARRIED	206021111P47230			[206021111P56414, 206021111P64725]	
206021111P42280	22	Male	MARRIED	206021111P47231			[206021111P54317]	
206021111P42281	24	Male	MARRIED	206021111P47232				
206021111P42282	24	Male	MARRIED	206021111P47233				
206021111P42283	24	Male	MARRIED	206021111P47234				
206021111P42284	16	Male	MARRIED	206021111P47235				
206021111P42285	18	Male	MARRIED	206021111P47236				

22,767 Rows, 9 columns, We will only be using 2 starting variables – Age and Gender

THE TASK

- Create a VISTA-like dataset from the Synthetic Population Dataset

  - Find correlations in the VISTA data
  - Find patterns of Trips and Activities based on basic factors like Age, Gender etc.
  - Use these patterns to predict data in the synthetic population

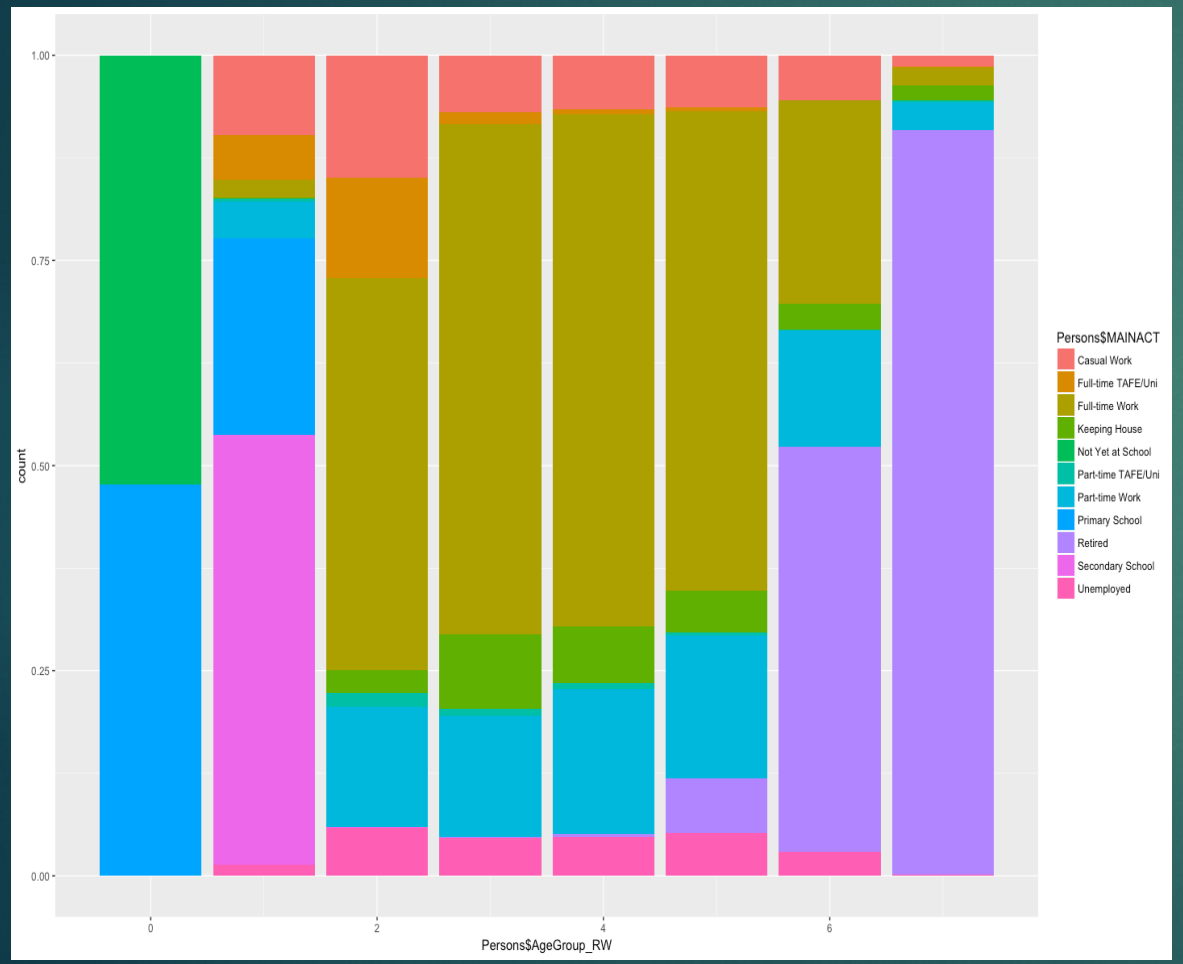
Main Activity

```
training_set <- subset(train_data,train_data$AgeGroup_RW==a)
test_set <- subset(test_data,test_data$AgeGroup==a)
test_set <- test_set[order(test_set$Rand),]
WorkCat<-as.data.frame(cumsum(wtd.table(training_set$MAINACT,weights =
training_set$CW_ADPERSWGT_LGA))/sum(wtd.table(training_set$MAINACT,weights = training_set$CW_ADPERSWGT_LGA))*100)
WorkCat$Act <- rownames(WorkCat)
rownames(WorkCat)<-c()
WorkCat$Wt <- WorkCat$`cumsum(wtd.table(training_set$MAINACT, weights =
training_set$CW_ADPERSWGT_LGA))/sum(wtd.table(training_set$MAINACT, weights = training_set$CW_ADPERSWGT_LGA)) * 100`
WorkCat <- WorkCat[,c(-1)]
j=1
i=1
for (i in 1:13){
  while (test_set$Rand[j]<WorkCat$Wt[i]&&!is.na(test_set$Rand[j])){
    test_set$MainAct[j] <- WorkCat$Act[i]
    j=j+1
  }
}

test_set<-test_set[order(test_set$AgentId),]
finalset <- rbind(finalset,test_set)
}
```

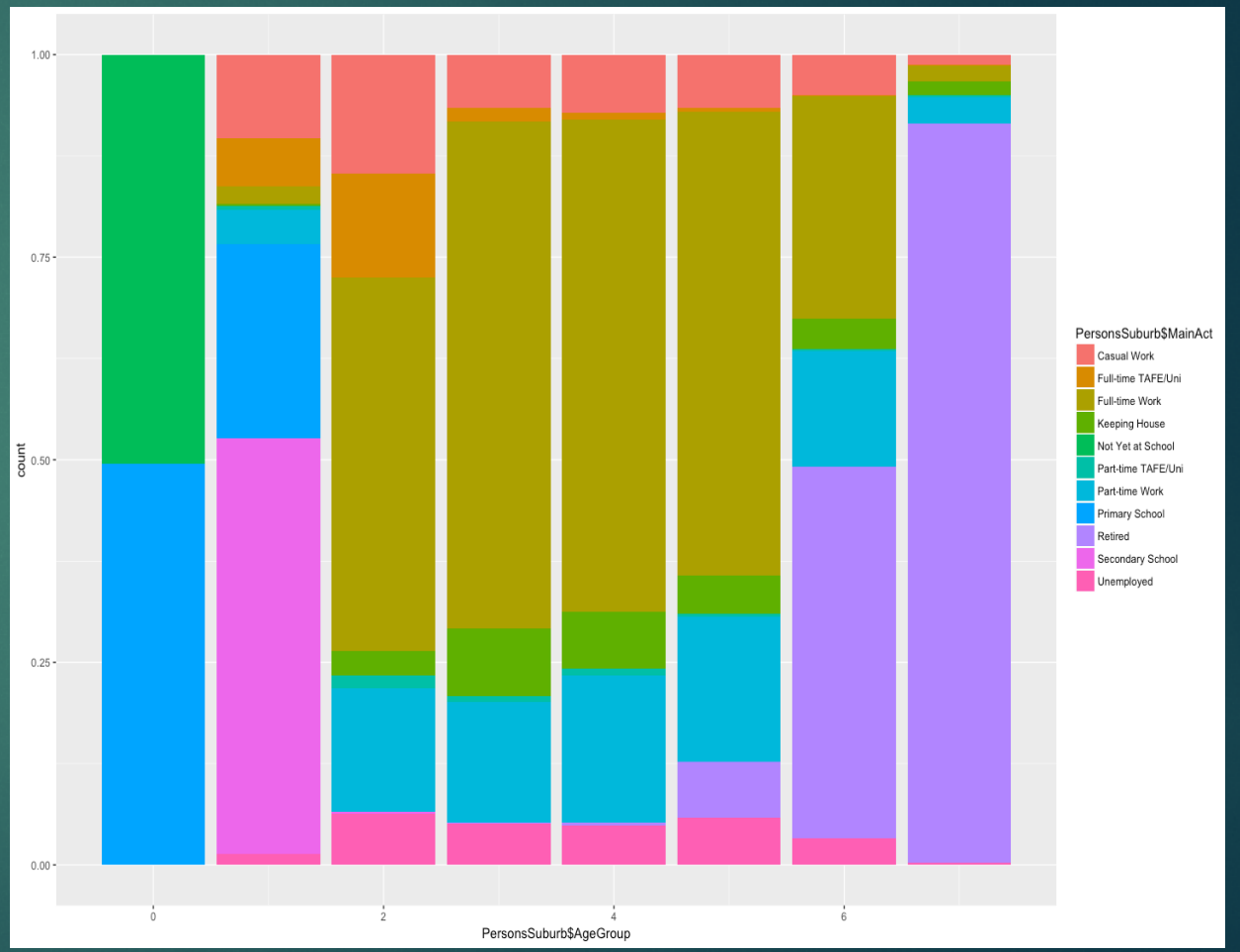
	Act	Wt
1	Casual Work	1.665255
2	Full-time TAFE/Uni	1.695194
3	Full-time Work	5.639915
4	Keeping House	5.796141
5	Not Yet at School	5.796141
6	Part-time TAFE/Uni	5.888355
7	Part-time Work	10.139719
8	Primary School	10.139719
9	Retired	99.790979
10	Secondary School	99.790979
11	Unemployed	100.000000

AgeGroup=7

Main Activity



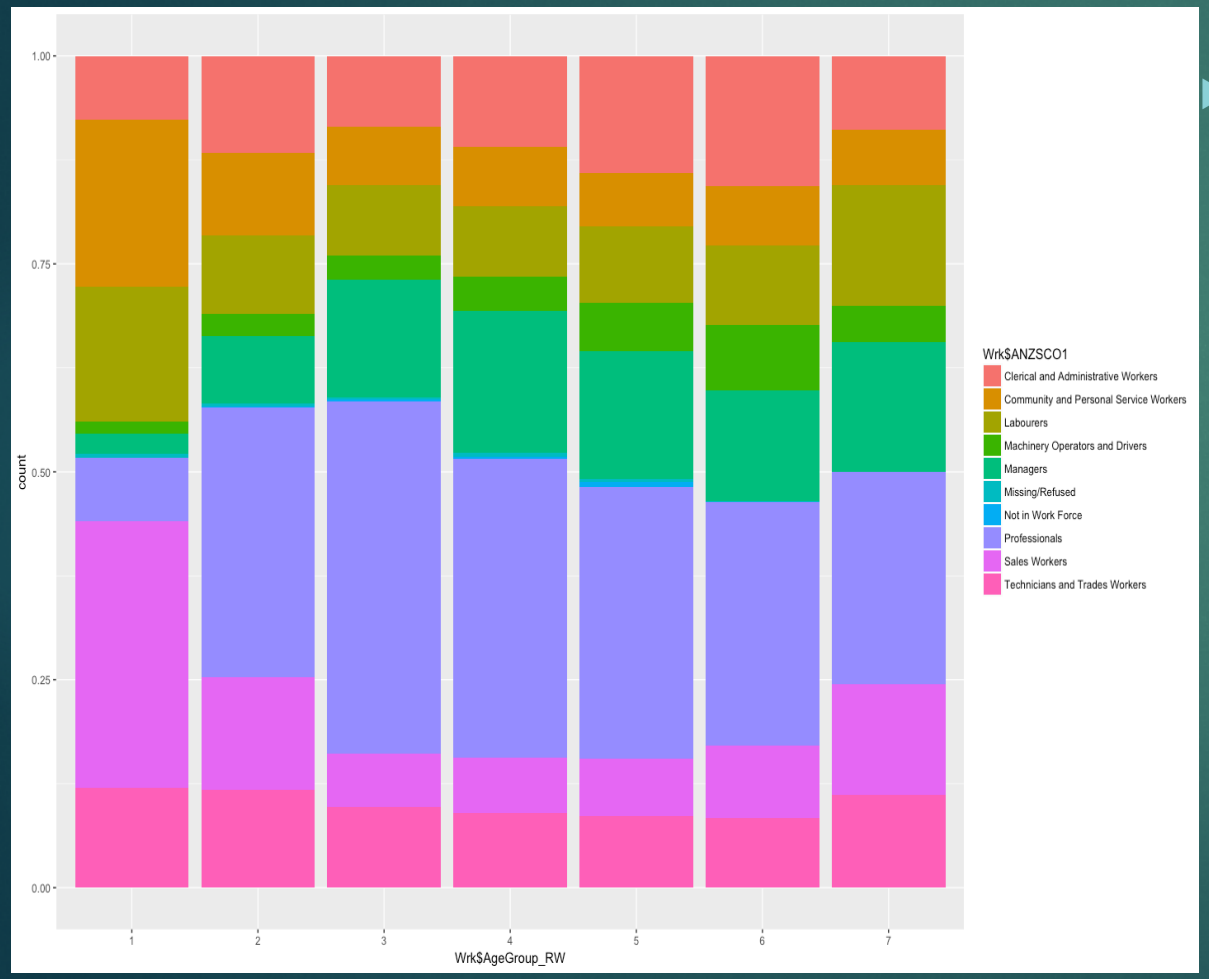
Known



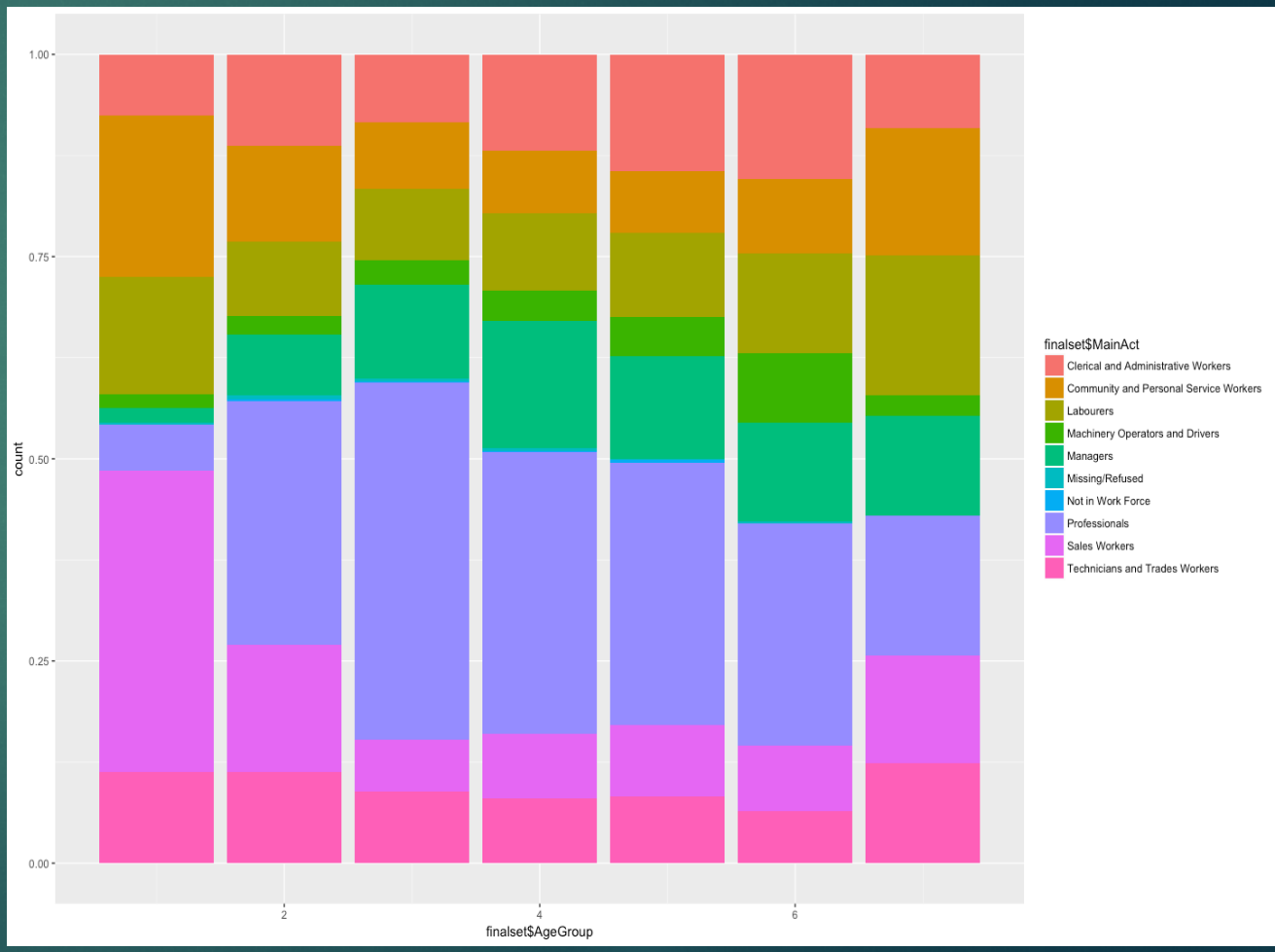
Unknown

ANSZCO – Main act + Age

Only Using Work Force (Casual Work, Partial Work, Full time Work)
Same Code as Main Activity

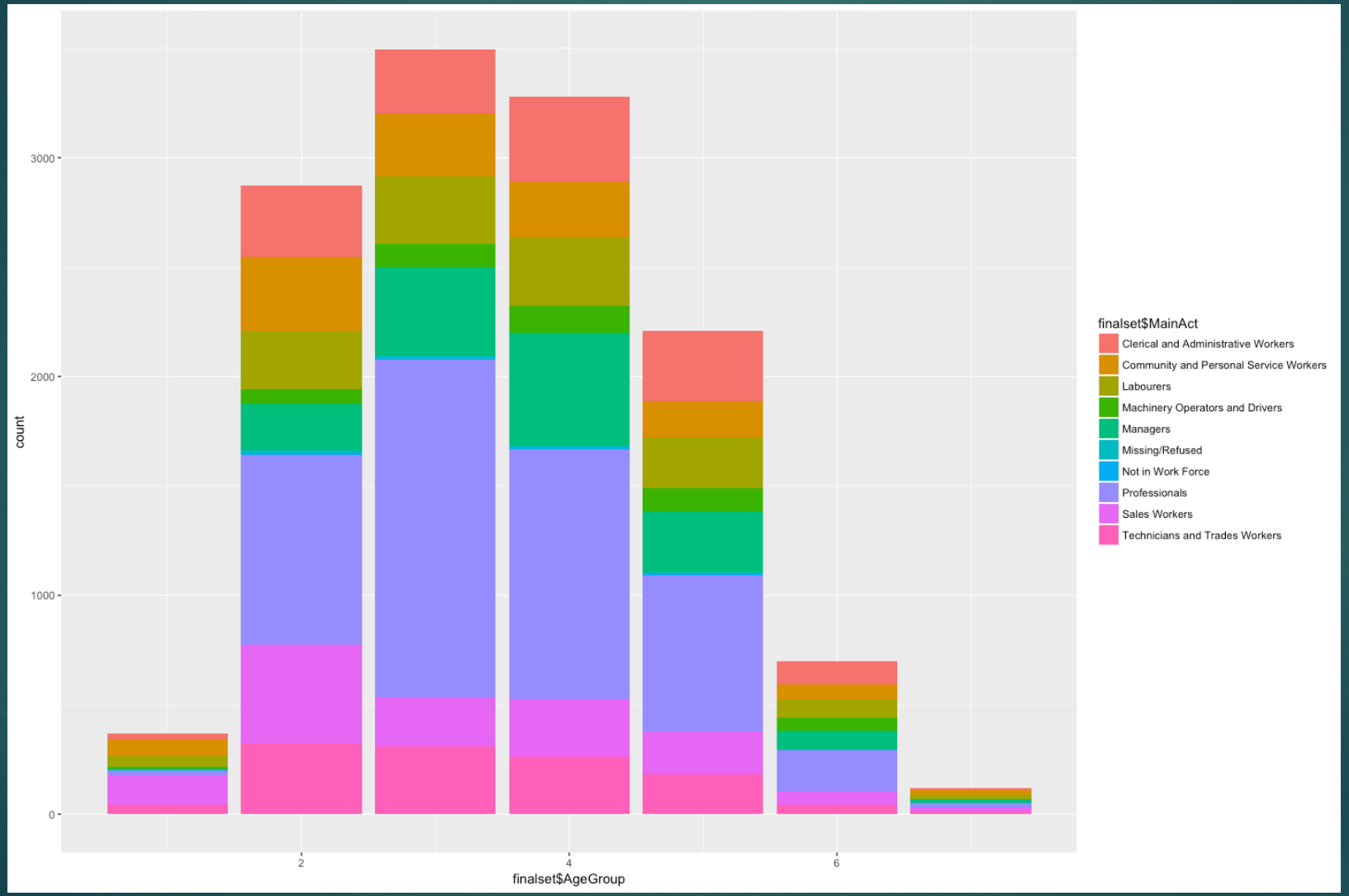


Known



Unknown

Work Dataset



Start time

```
Call:
lm(formula = Wrk$STARTIME ~ Wrk$SEX + Wrk$AGE + Wrk$MAINACT)

Residuals:
    Min      1Q  Median      3Q     Max
-356.92  -74.80  -21.95   27.34 1202.18

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 619.2507     6.4401  96.156  < 2e-16 ***
Wrk$SEXMale                  -8.2794     2.9021  -2.853  0.00434 **
Wrk$AGE                      -0.8934     0.1090  -8.195 2.78e-16 ***
Wrk$MAINACTFull-time Work  -109.7859     5.3375 -20.569  < 2e-16 ***
Wrk$MAINACTPart-time Work   -39.2046     6.1243  -6.401 1.60e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 143.8 on 10903 degrees of freedom
Multiple R-squared:  0.07274,   Adjusted R-squared:  0.0724
F-statistic: 213.8 on 4 and 10903 DF,  p-value: < 2.2e-16
```

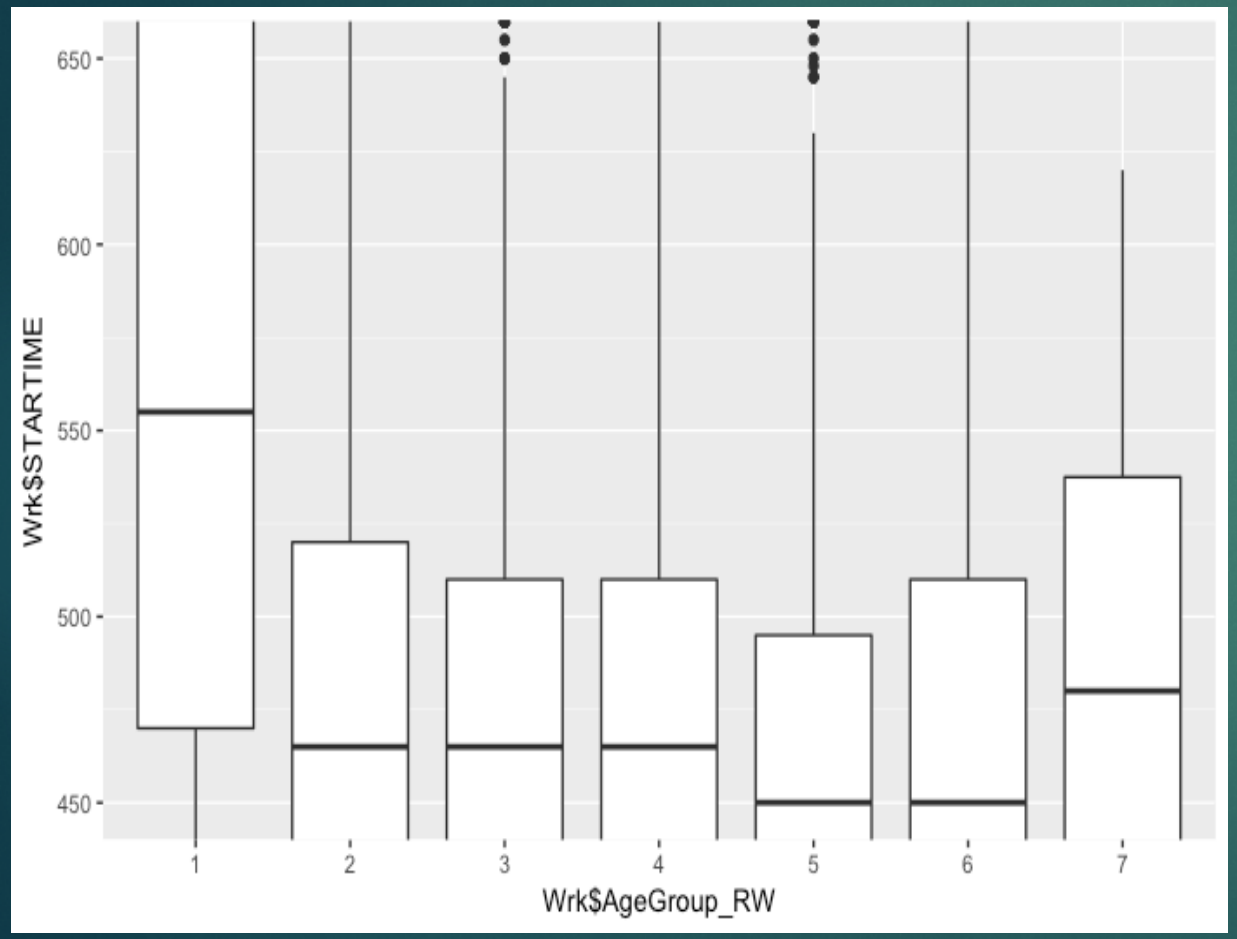
- Finding the variables with which Start Time has a high correlation with.
- Low R values signify that the variance will not be well represented
- Very low P-value signifies that the variables depend on each other a lot

Start time
Dependent on Age, Sex, MainAct

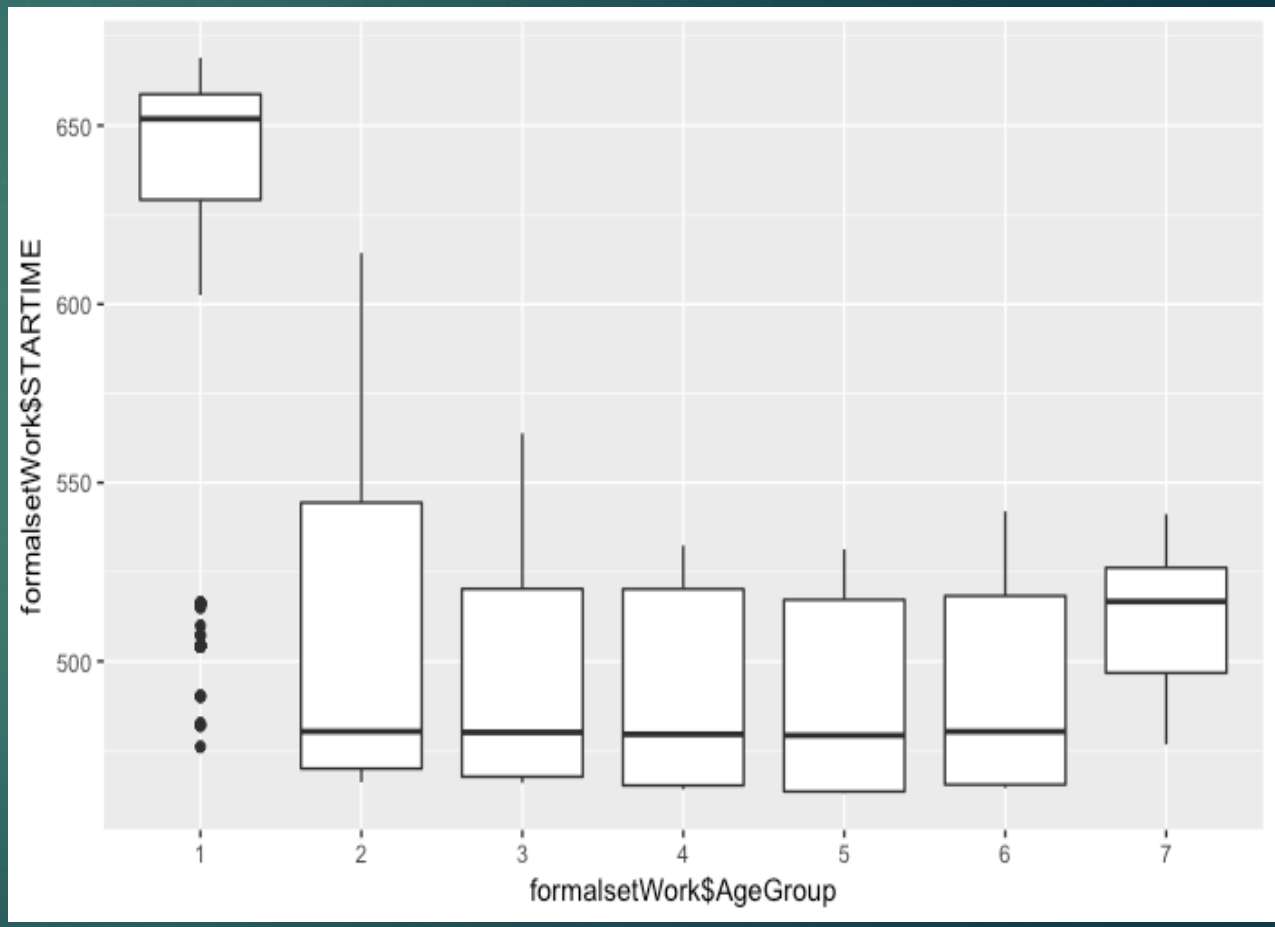
```
dataset2 <- Wrk[,c(114,116,128,11)]
dataset2$SEX = factor(dataset2$SEX,
                      labels = c(1, 2))
dataset2$MAINACT = factor(dataset2$MAINACT, labels = c(1:nlevels(dataset2$MAINACT)))
test_set <- formalsetWork[,c(2,3,13)]
colnames(test_set)[3]="MAINACT"

test_set$SEX = factor(test_set$SEX,
                      labels = c(1, 2))
test_set$MAINACT = factor(test_set$MAINACT, labels = c(1:nlevels(dataset2$MAINACT)))
line_reg = randomForest(x=dataset2[-4],y=dataset2$STARTIME,ntree = 100,weights = Wrk$CW_ADJTWWGT_LGA)
formalsetWork$STARTIME = predict(line_reg, test_set)
```

- Code for the RandomForest regression technique
- Trees can be an arbit number. Hit and trial (inside a loop) found 100 to be a decent representation

Start Time



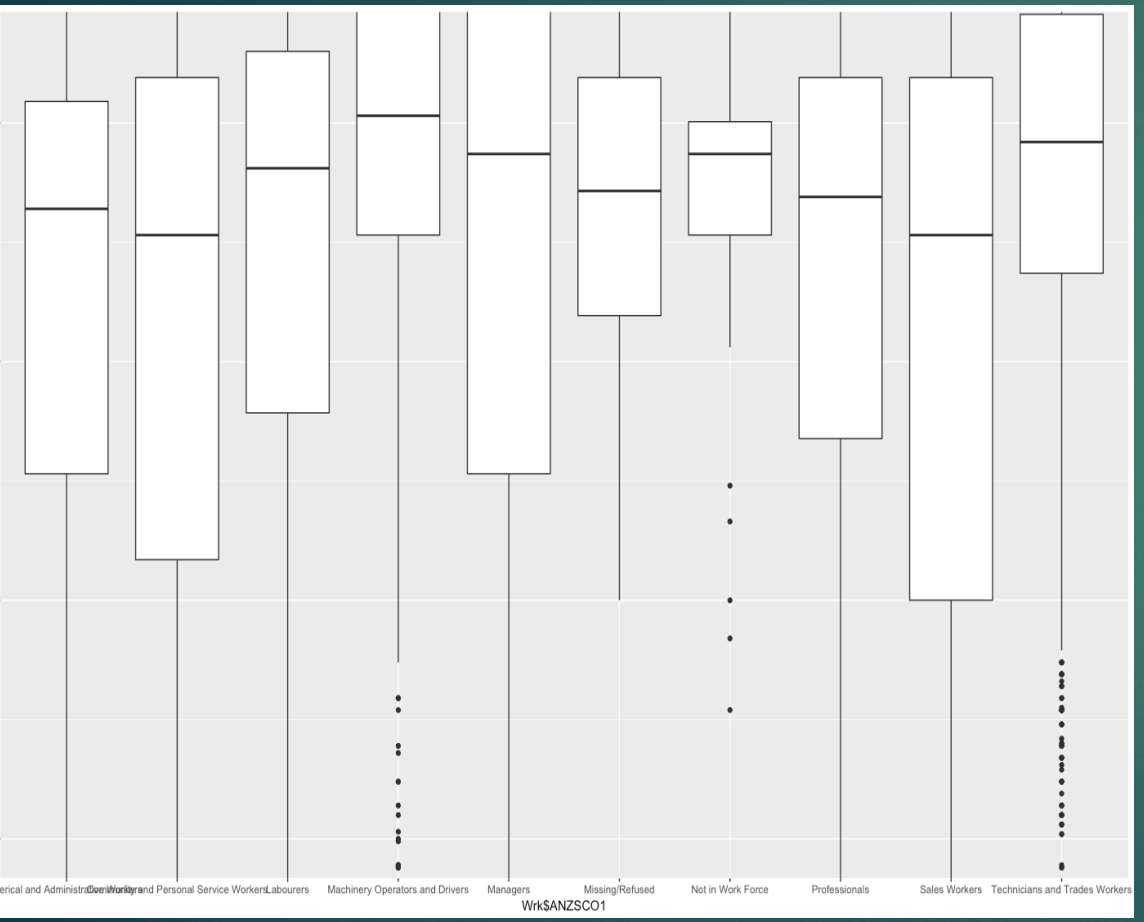
Known



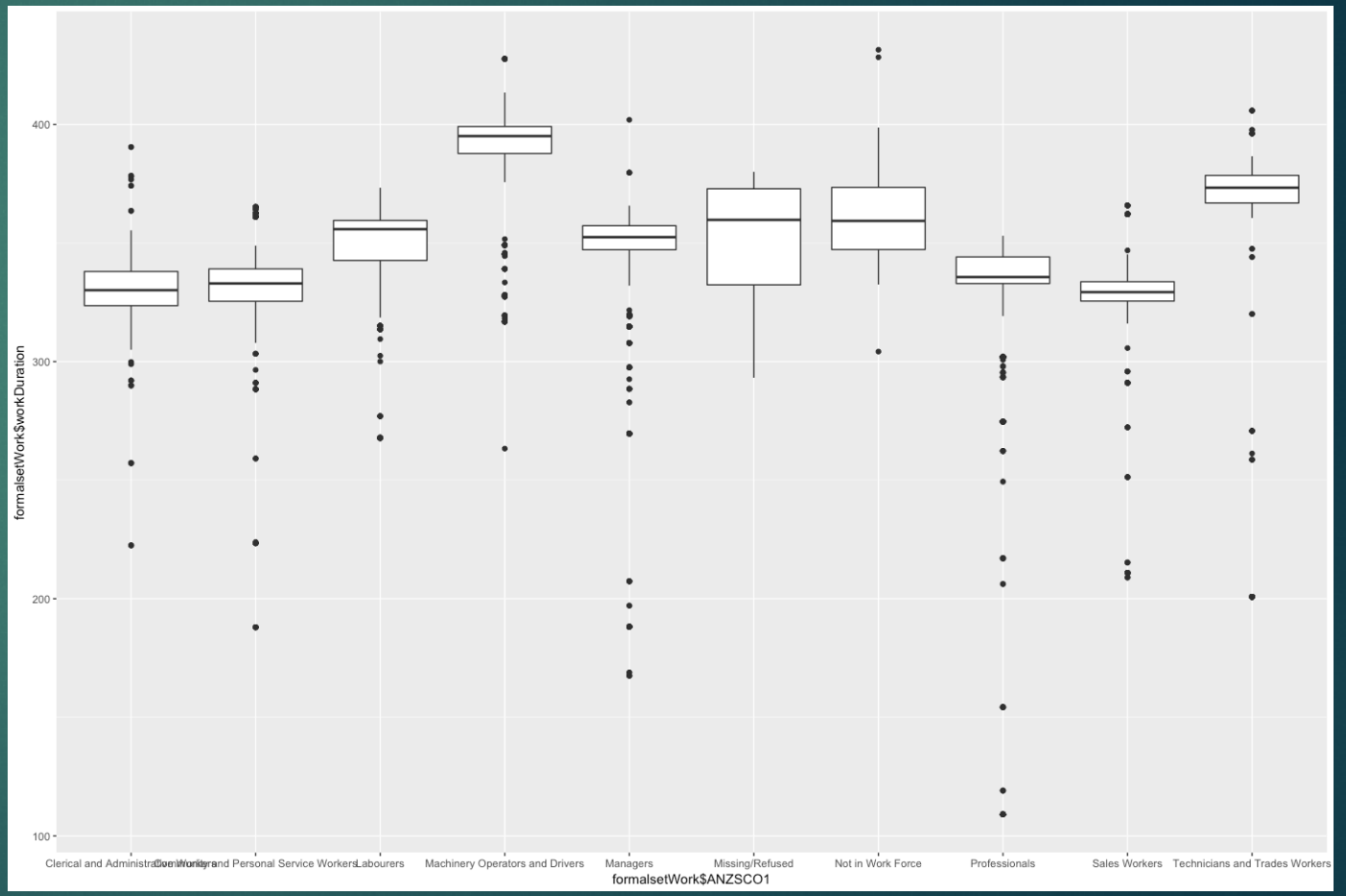
Unknown

Duration – Age, Sex, ANZSCO1

Graph made betweeen Duration and ANZSCO1



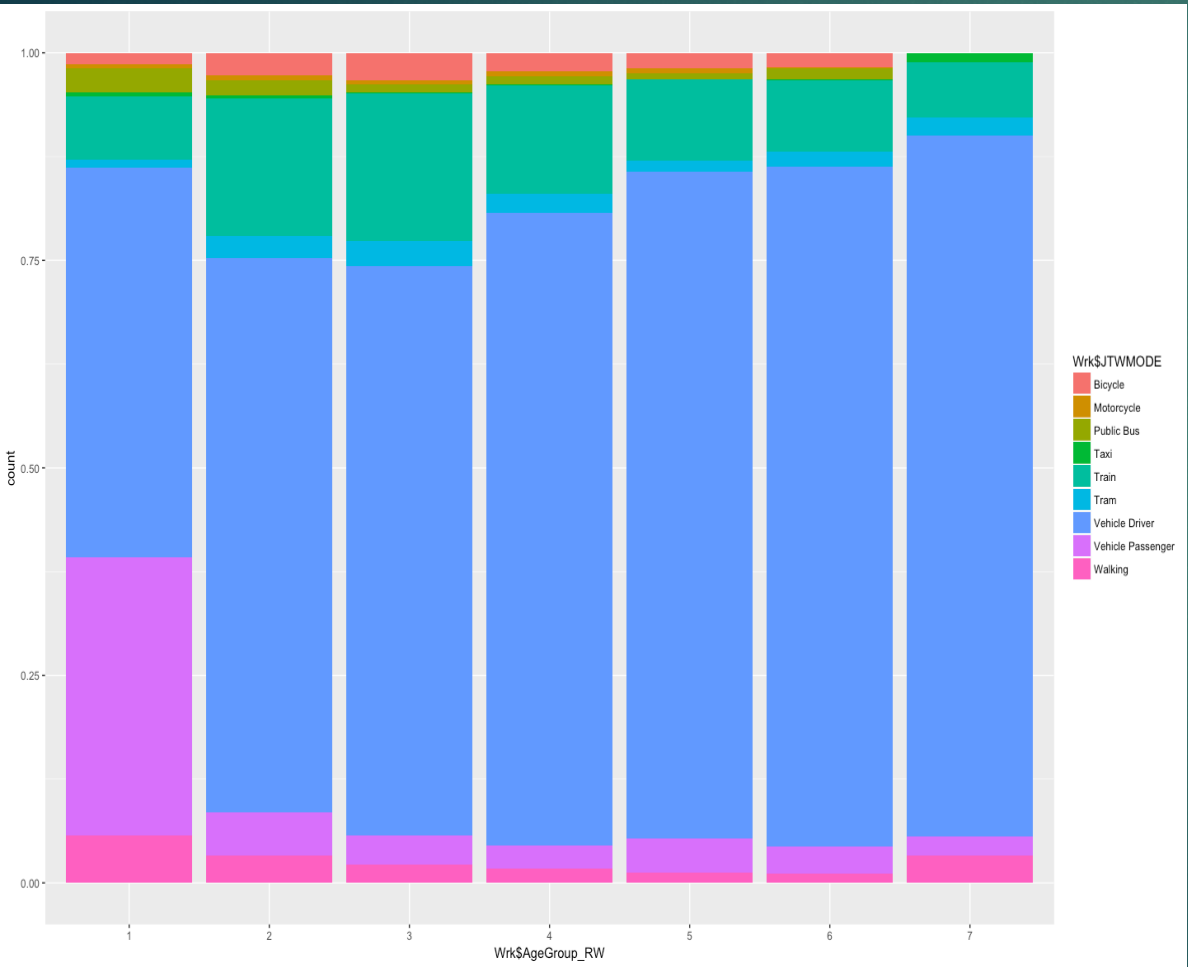
Known



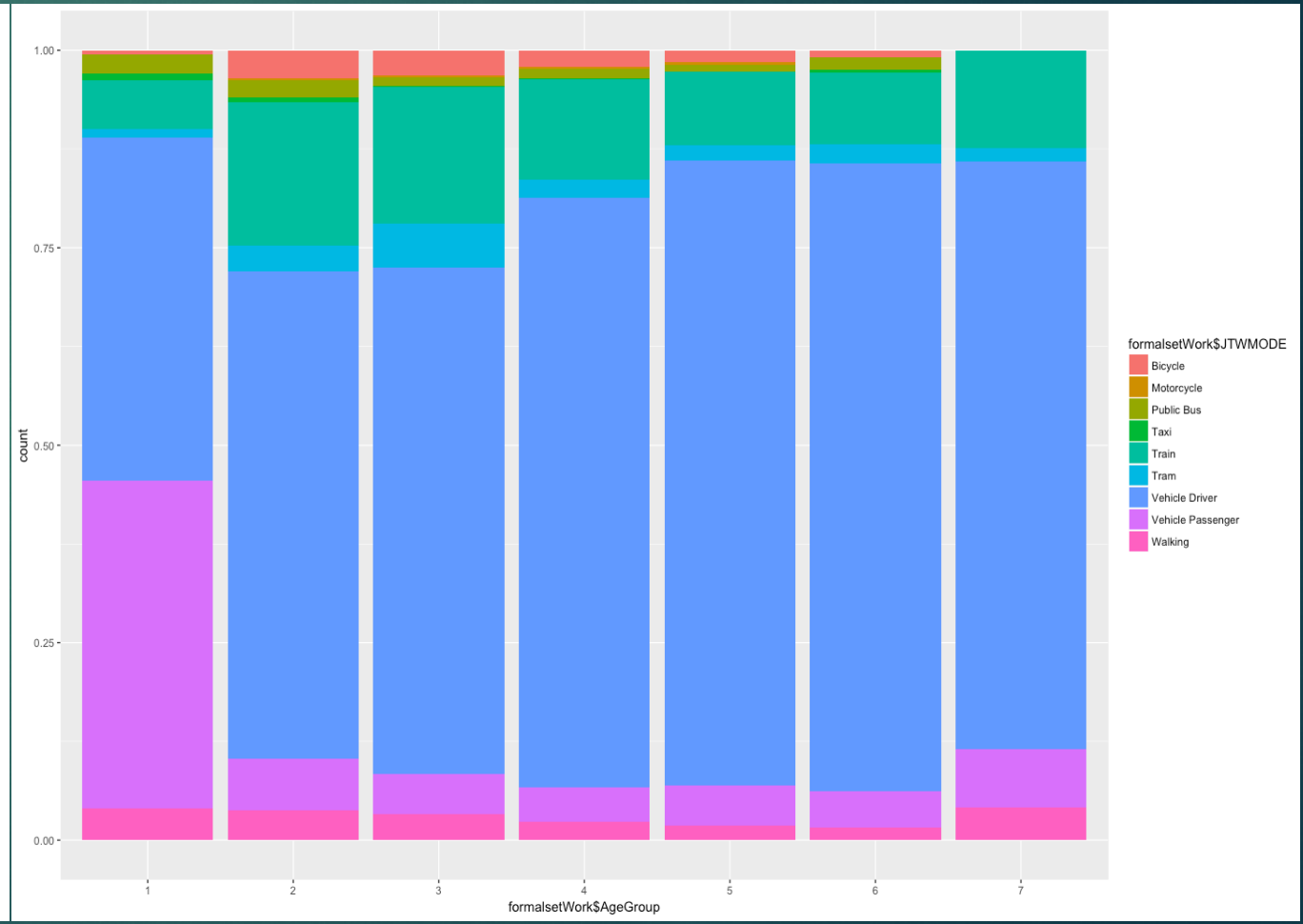
Unknown

Mode

Made using same method as Main Act and ANZSCO1



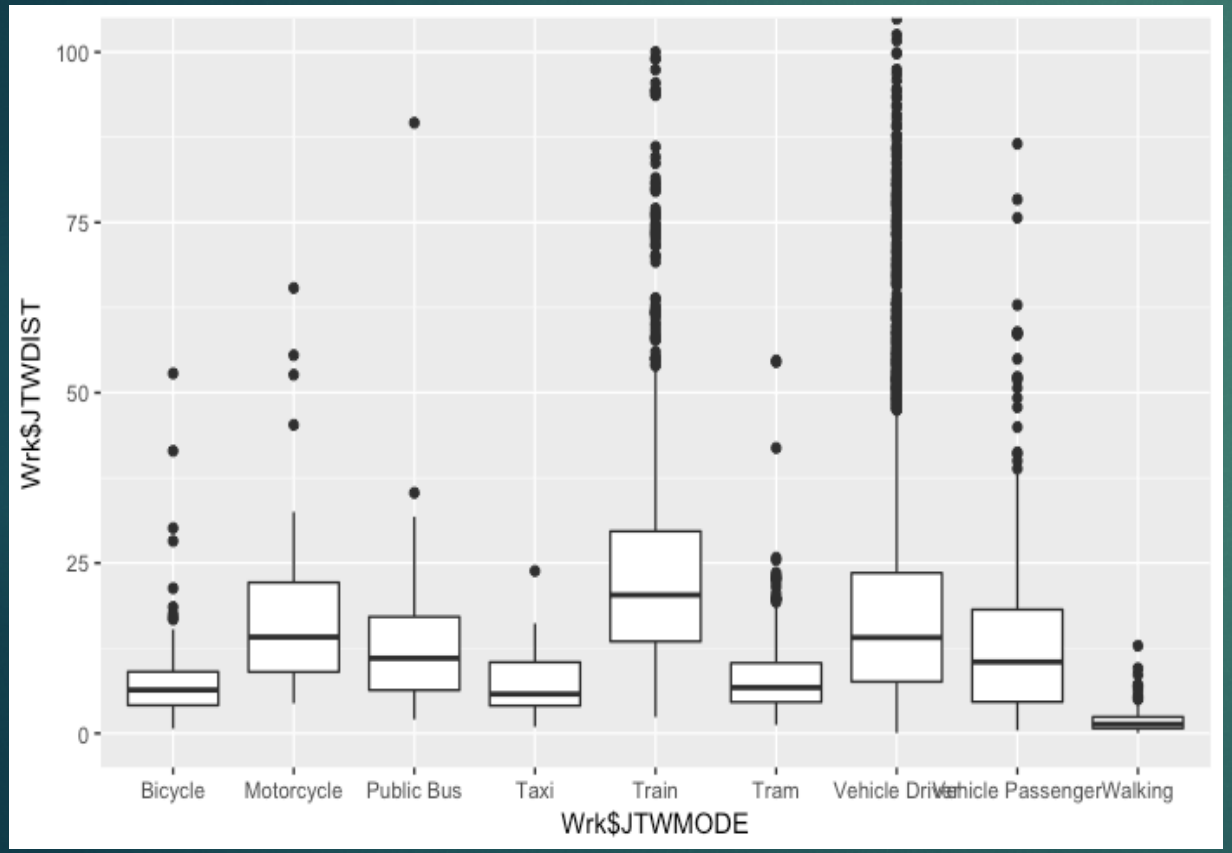
Known



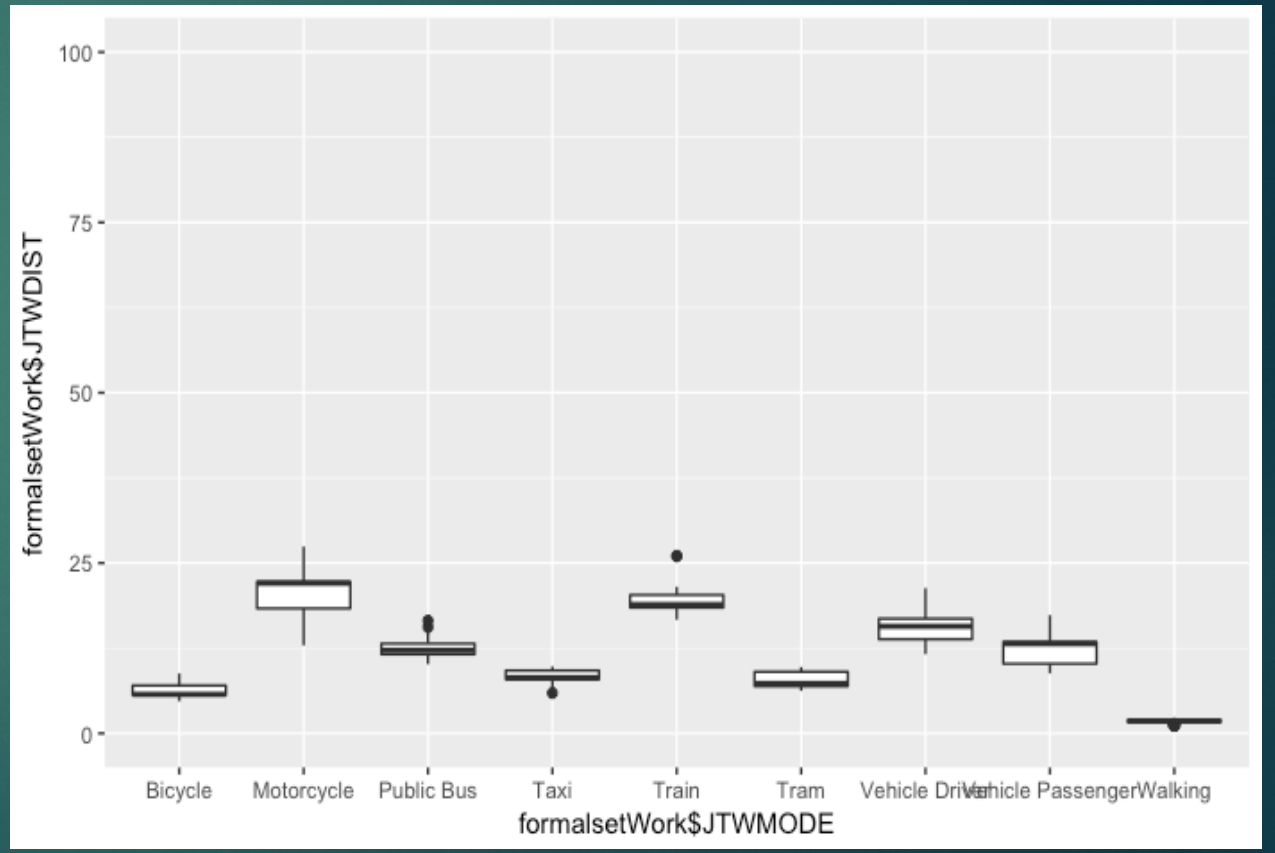
Unknown

Distance to Work

Dependent on Mode and Age

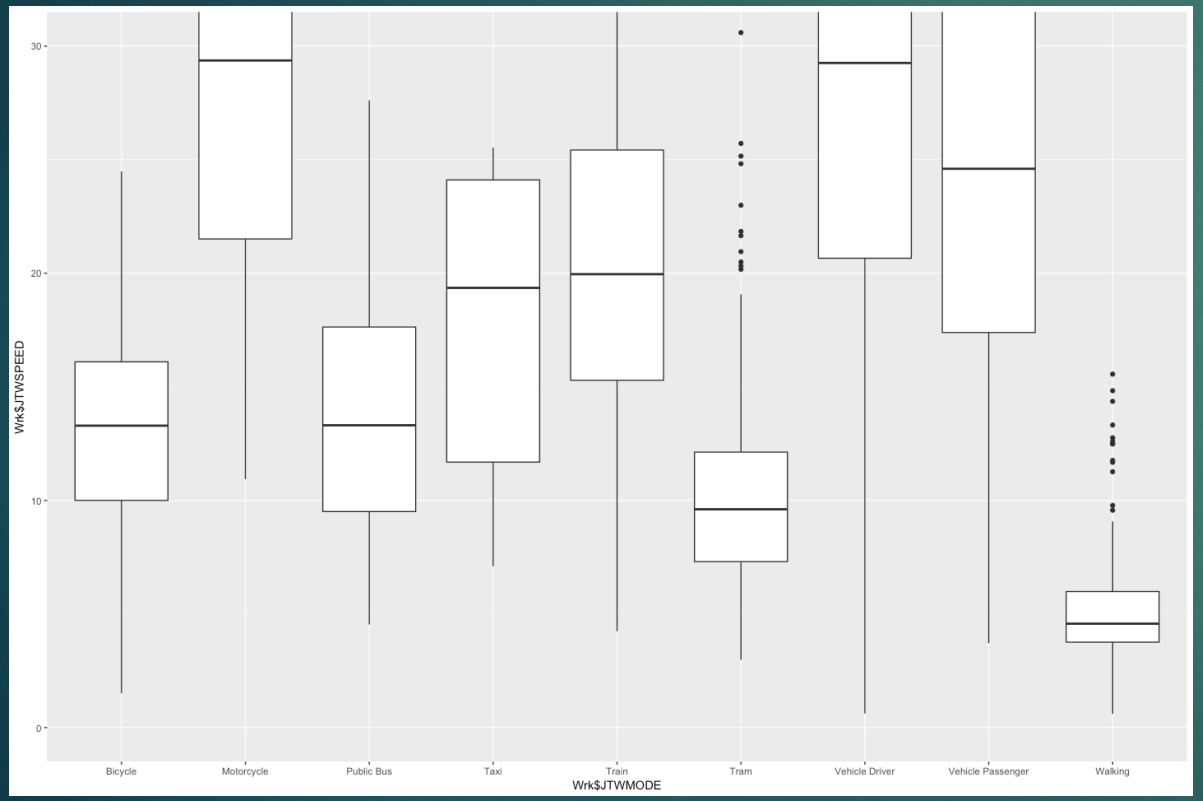


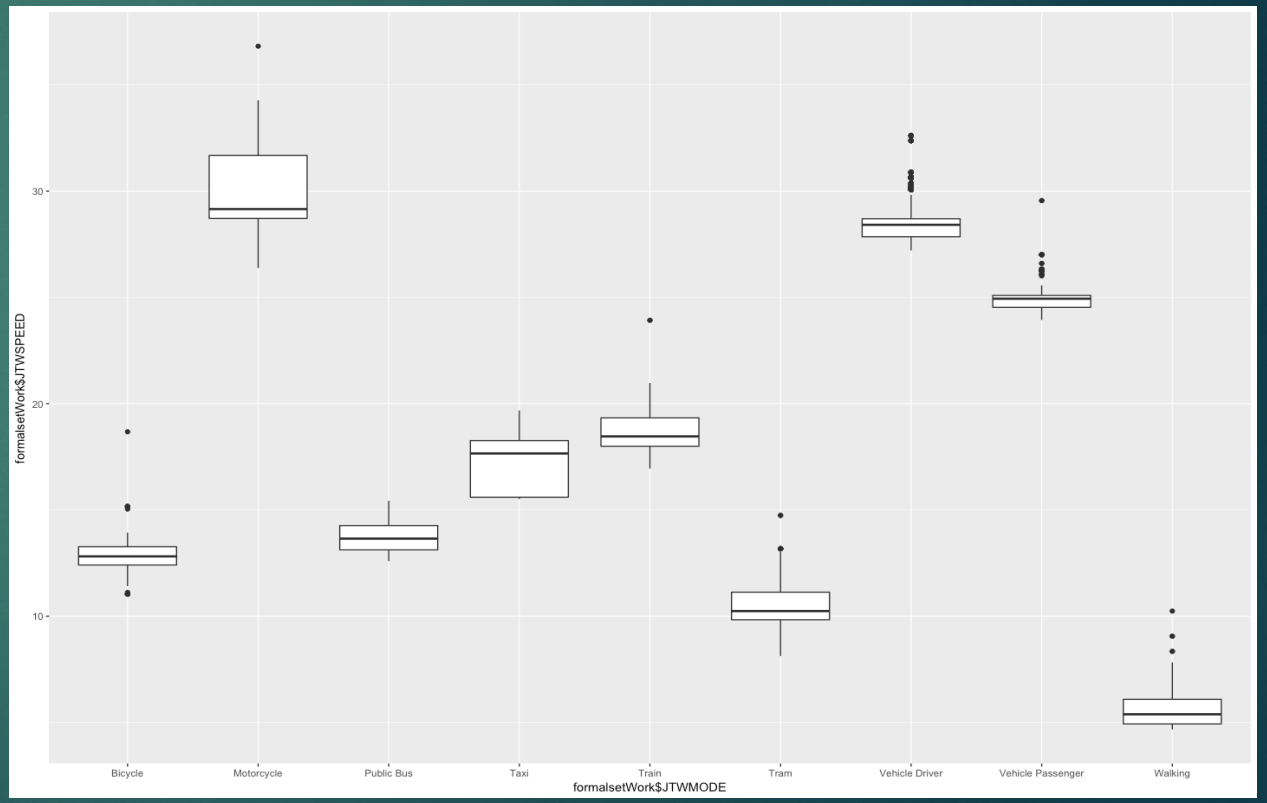
Known



Unknown

Ave Speed
Depends on Mode and StartTime





Known

Unknown

Calculations to complete trip

- We know speed, distance so we can get time of travel
- With simple addition, we now have:-
  - Time of reaching work
  - Duration of Work
  - Time of leaving work
- For Return trip, we assume same mode of travel. Distance remains the same.
- Same as finding Speed to work, we find speed from work
- Same calculations to get time required to reach home.

- Considering – Direct Journey from work to home.
- 10000 out of 16000 trips from work are directly
- Talking about Accompany Someone, Buy Something, Change Mode etc. these are small stops which do not affect the journey that much.
- We need to analyze correlation between different Activities.

Accompany Someone	At or Go Home	Buy Something	Change Mode	Education
44	10259	1145	18	31
Not Stated	Other Purpose	Personal Business	Pick-up or Deliver Something	Pick-up or Drop-off Someone
1	18	289	142	745
Recreational	Social	Work Related		
367	1114	2277		

Education

- Whatever was done for Work, we use the exact same steps for Education.
- We hence find Start time, travel time, duration of Education, departure time and time of reaching home again.
- Because timings of school/college are more fixed than work timings, errors are less in this data. (Both variances are lesser than that of Work)
- We can merge the two sets (work, education) to form one set of everyone who is either in the work force or a student.
- Because the lifestyle of these people, will be different from others (Infants, Retired etc.) we will always keep the two datasets different.

People not in the Work Force or Students (NonWork)

- There are majorly 4 categories of people left to be analysed
  - Retired
  - Unemployed
  - Homemakers
  - Infants
- We use a similar process that we used for Work Force
  - Giving each person a Major Activity of that day.
  - Giving him/her a Mode of travel
  - Predicting duration of activity, travel times etc.
- We then merge these 4 datasets together to form a grand "NonWork" dataset

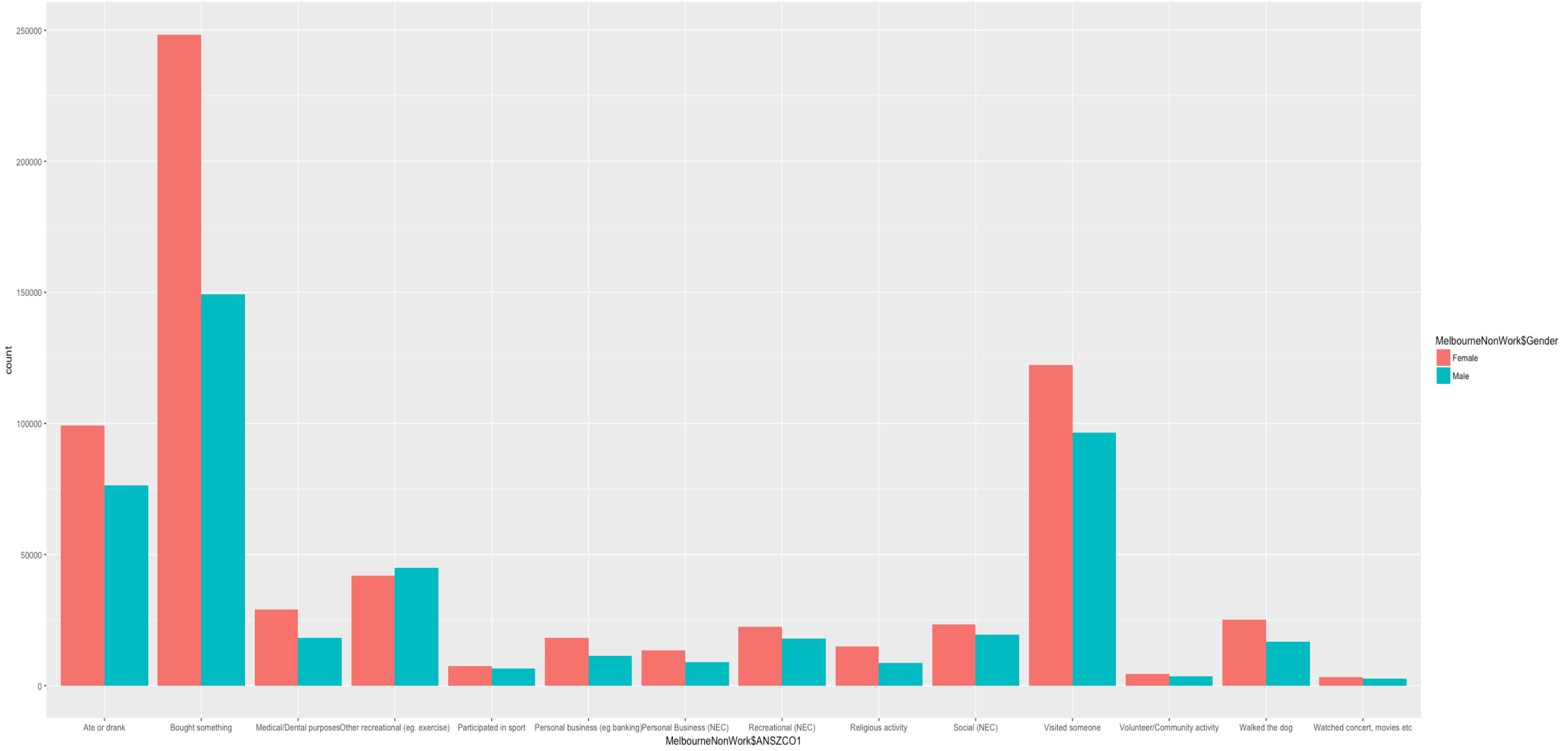
Final Datasets

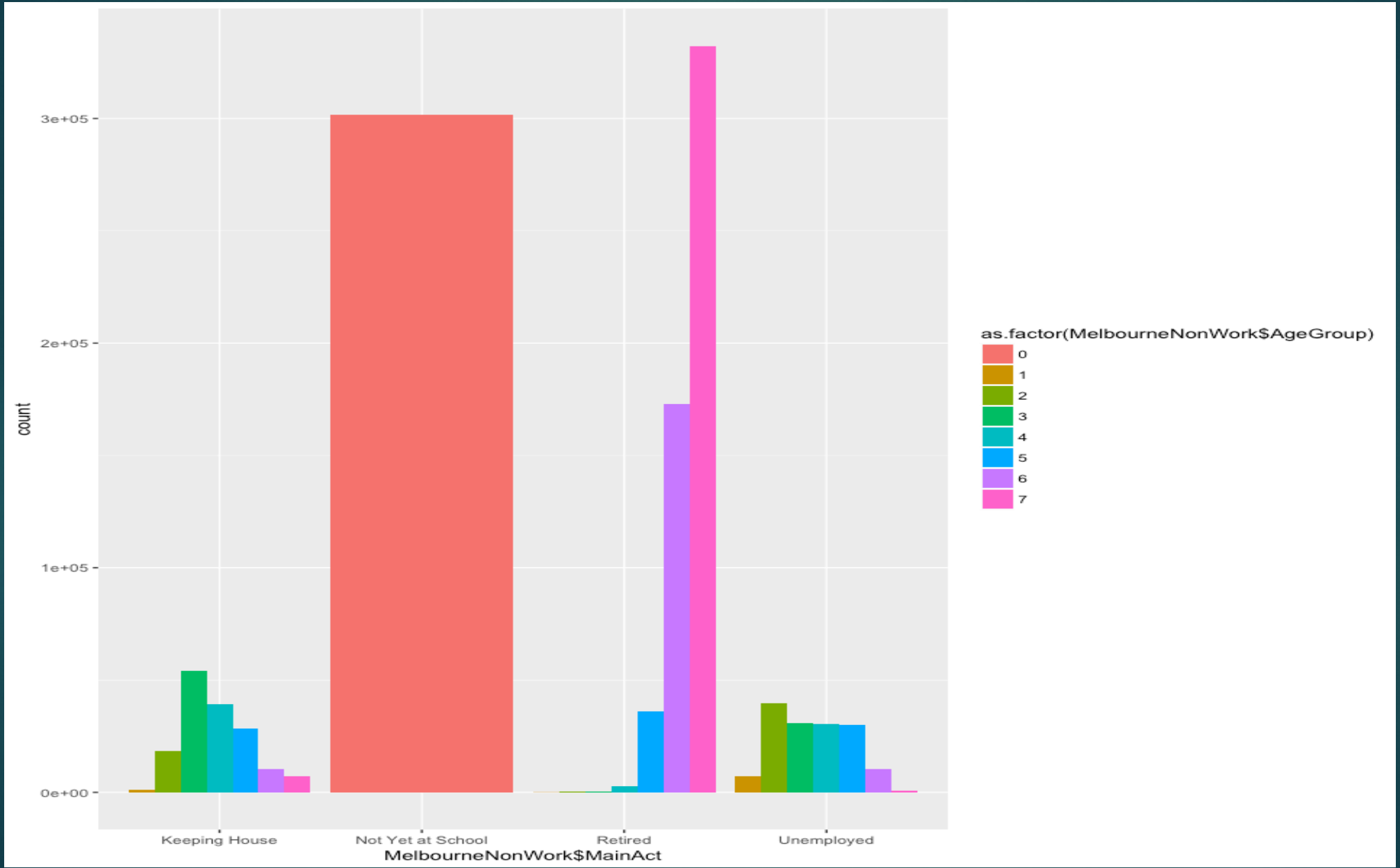
- We use a loop to access all directories and bind the data of all the suburbs.
- Individually, we use the same methods on each dataset to get information of trips.
- Then we bind this final set of all suburbs to get two grand datasets for Melbourne
  - Work
  - NonWork

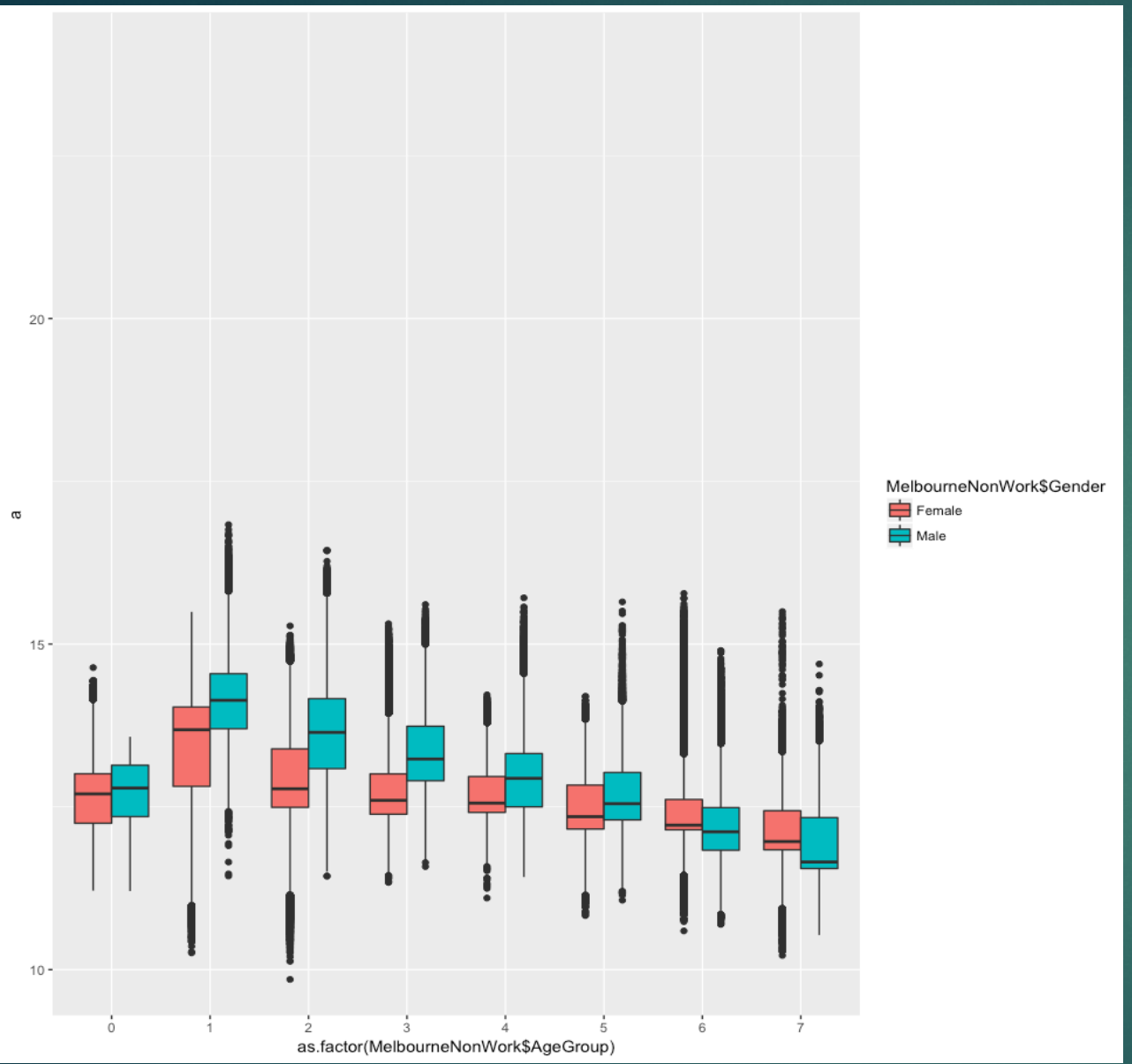
- Melbourne Non Work - 1.16 Million entries
- Melbourne Work – 3.1 Million entries

- We have the entire population of Melbourne, and total information of one activity that they have done in that day

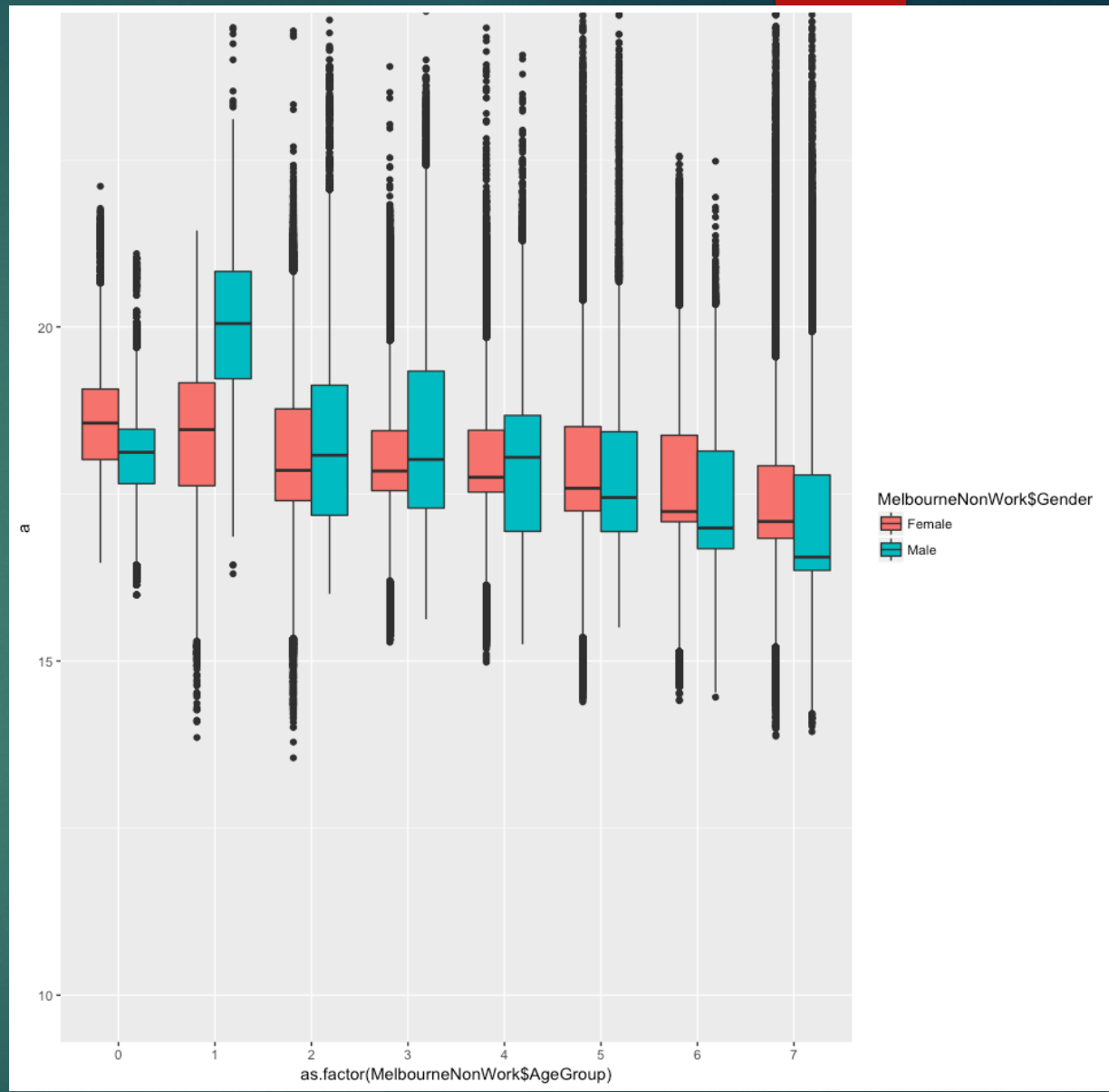
Some Graphs


250000
200000
150000
100000
50000
0
count
Ate or drank
Bought something
Medical/Dental purposes
Other recreational (eg. exercise)
Participated in sport
Personal business (eg banking)
Personal Business (NEC)
Recreational (NEC)
Religious activity
Social (NEC)
Visited someone
Volunteer/Community activity
Walked the dog
Watched concert, movies etc
MelbourneNonWork$ANSZCO1
MelbourneNonWork$Gender
Female
Male

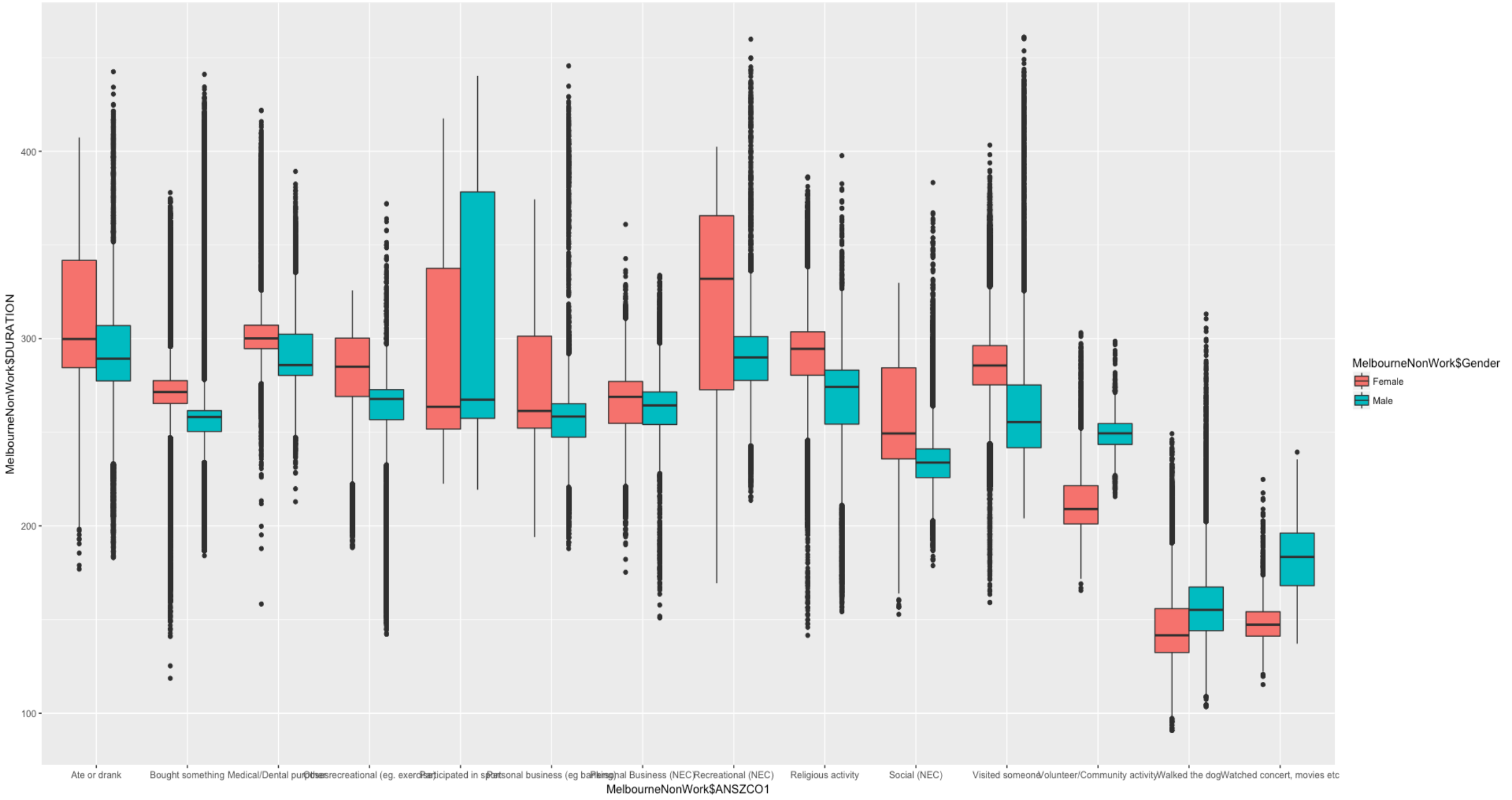

3e+05
2e+05
1e+05
0e+00
count
Keeping House
Not Yet at School
Retired
Unemployed
MelbourneNonWork$MainAct
as.factor(MelbourneNonWork$AgeGroup)
0
1
2
3
4
5
6
7



Start Time



Reach Home Time


MelbourneNonWork$DURATION
400
300
200
100
Ate or drank
Bought something
Medical/Dental purposes
Other recreational (eg. exercise)
Participated in sport
Personal business (eg banking)
Personal Business (NEC)
Recreational (NEC)
Religious activity
Social (NEC)
Visited someone
Volunteer/Community activity
Walked the dog
Watched concert, movies etc
MelbourneNonWork$ANSZCO1
MelbourneNonWork$Gender
Female
Male

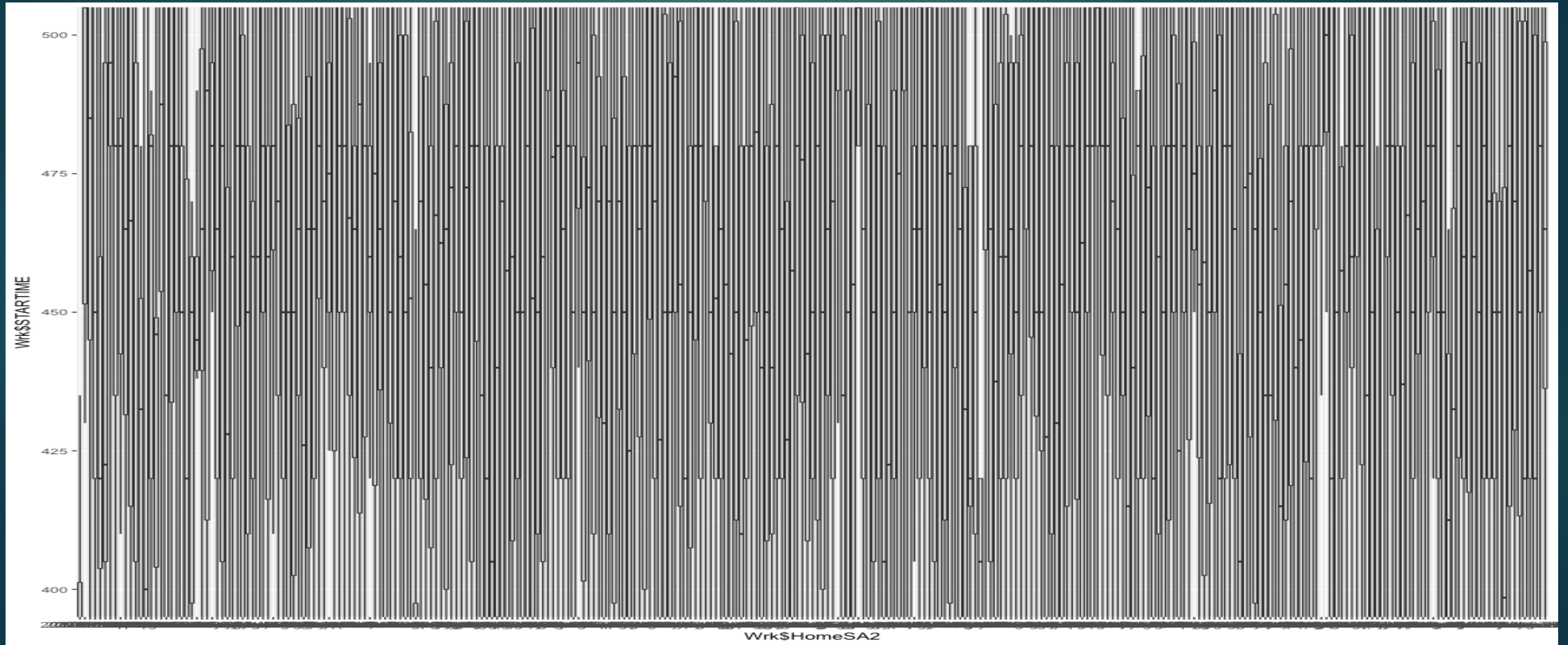
Precautions

- Factors with less frequency may not show up at all in suburb set. Need to keep sub-setting train set according to test set to avoid errors in regression
- Balancing the fine line between precision and accuracy is tough. Do we want our data to be accurate to the train data's variance? Or do we want precise results? Currently, we aim for precision.

Error Handling

- Find a regression model that correctly predicts the mean and the variance of the training data
- Observe differences between suburbs (need a larger dataset per suburb to use it as a training data set)
- Using smaller datasets (per suburb) which will (hopefully) reduce variance.

Limitations – Correlation between Suburbs of Melbourne



Limitations – Correlation between Suburbs of Melbourne

- Although there are differences in each suburb, it is tough to address them because :-
  - We don't have much data about each particular suburb and hence training set is very poor
  - We have not used geographical maps to find out which suburbs are near each other (For example, Yarra North and Yarra Central will have similar patterns, as compared to Mount Eliza)
- So we have used a training data of the entire city for each suburb.
- To correct this, in depth analysis of many factors like Tram Stops, Train stations, freeways etc. will have to be done.
- Till then, each person in Melbourne is considered equal.

Limitations – Correlations between Different People

Accompany Someone	479698.74
At or Go Home	5386973.70
Buy Something	1436712.54
Change Mode	13409.15
Education	568899.19
Not Stated	222.74
Other Purpose	49043.52
Personal Business	723456.60
Pick-up or Deliver Something	183818.97
Pick-up or Drop-off Someone	957121.60
Recreational	720737.53
Social	1443339.16
Work Related	1704945.51

"Accompany Someone" and "Pic—up or Drop-off Someone" are two categories which are tough to analyze.
Ex. Mode of trip – Vehicle Passenger. We do not know who was the driver of the vehicle.

In the prediction, we have predicted each Agent independently of his/her family.
Ex. We have information of all trips to Work. We do not know whether there was a stop to drop off child at school. However, we have information of that child's trip to school.

Limitations – Day of Survey vs Average Day

- We just get the trips taken by the person on that particular day. That may not be characteristic to the trips that person does on an average day.
- For example, if a particular suburb has more visits with a purpose of religious activity, it may be because there are more churches, or it may be because the survey was done on a Sunday.
- Religious Activity (Sunday) – 51030.01
- Religious Activity (NotSunday) – 4157.723
- As we are finding trips on an average day, we ignore effects of particular days and use the same algorithm for every day. We also assume that the VISTA data is for an average day of the entire population.

Limitations – Correlation between Activities

- Depending on type of work, age, distance to work etc. a person may have different activities.
- Depends on timings of school and work, the free time of people is also different. Morning activities are different from Evening activites.
- Watching movies (evening) – 35000 trips
- Watching movies (morning) – 8000 trips
- This problem becomes more relevant to people not in work force because they have the entire day to do different activities

Work Related	Work Related	222462.32
Buy Something	Buy Something	178612.24
Social	Social	144726.85
Social	Work Related	120082.31
Buy Something	Work Related	116062.35
Buy Something	Social	113010.15
Work Related	Pick-up or Drop-off Someone	102532.89
Social	Buy Something	95607.67
Work Related	Social	92948.90
Pick-up or Drop-off Someone	Pick-up or Drop-off Someone	91511.04
Pick-up or Drop-off Someone	Work Related	76229.46
Accompany Someone	Accompany Someone	72130.35
Buy Something	Personal Business	70805.24
Work Related	Buy Something	66999.96
Buy Something	Pick-up or Drop-off Someone	61325.94
Social	Pick-up or Drop-off Someone	60735.72
Personal Business	Social	54238.18
Social	Recreational	49836.98
Pick-up or Drop-off Someone	Social	49774.55
Social	Personal Business	43935.07

High Correlation

Education	Buy Something	4036.54
Education	Recreational	3840.85
Pick-up or Deliver Something	Education	3436.39
Pick-up or Deliver Something	Recreational	3291.50
Education	Work Related	3244.49
Pick-up or Drop-off Someone	Unknown purpose (at start of day)	3116.84
Education	Personal Business	2715.90
Accompany Someone	Unknown purpose (at start of day)	2685.35
Recreational	Pick-up or Deliver Something	2633.78
Education	Pick-up or Deliver Something	1537.78
Pick-up or Deliver Something	Unknown purpose (at start of day)	1242.55
Pick-up or Deliver Something	Accompany Someone	1234.44
Accompany Someone	Pick-up or Deliver Something	734.28

Low Correlation

Next Step?

- Give more Activities to people. Analyze what time a person is free and likelihood to do an activity then
- Find effect of Household on people. Dependencies amongst people in a family
- Use geographical data to find factors that increase/decrease likeliness of particular modes of travels and travel times/distances.

Alternative Approach?

- Our entire model is built per person. We have seen the characteristics of the person and given him/her certain activities and trips
- The issue here is deciding the number of activities to be given and then spacing these activities out.

- We choose trips based on age and gender of the population (and geographical factors if found) and then we also use household data to assign trips to a particular household.
- This takes care of correlations between activities (correlations exist in training set), and correlations among people (trips are assigned to household not people).