THE UNIVERSITY OF TEXAS AT DALLAS

INTELLIGENCE ANALYTICS SOCIETY

CHALLENGE 3.0

HELPING THE WORLD, ONE DATASET AT A TIME

TEAM
RANDOM

II.

HOW DOES TIME SPENT ON THE VARIOUS ACTIVITIES REFLECT THE EMPLOYMENT STATUS OF AN INDIVIDUAL?

TEAM RANDOM

USING 4 DIFFERENT MODELS, WE CAN ACHIEVE A PREDICTIVE ACCURACY OF 67 - 85%*

PROJECT OUTLINE

THREE SMALL STEPS WE TOOK TO MAKING AN AWESOME MODEL...

DATA CLEANING

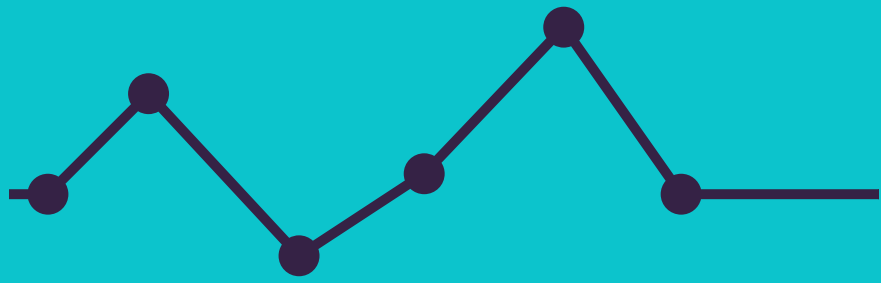
OUTLIER REMOVAL; OVERLAPPING CORRECTION

DATA VISUALIZATIONS

MACHINE LEARNING

KNN-CLASSIFIER; SUPPORT VECTOR MACHINE; RANDOM FOREST; C50 DECISION TREE

PROJECT ROADMAP



TRAINING DATASET

OUTLIER REMOVAL

WITH CORRECTION

WITHOUT CORRECTION

DATA VISUALIZATION

MACHINE LEARNING

TEST DATASET

OUTLIER REMOVAL

WITH CORRECTION

WITHOUT CORRECTION

PREDICTION

OUTLIER REMOVAL

THE EXTREME OBSERVATIONS THAT DOESN'T MAKE SENSE ON AN AVERAGE BASIS

Does it make sense for an employed individual to have no weekly hours but still earn?

NO

Or.. does it make sense for an employed individual to work but not have any weekly earnings?

NO

How about a person who spends more than 14 hours sleeping on an average basis?

POSSIBLE.. BUT HIGHLY UNLIKELY

II.

USING THE 2ND AND 98TH PERCENTILE AS A THRESHOLD, WE REMOVED ALL OUTLIERS...

WE EVEN IMPOSED A MINIMUM WAGE CRITERIA

LEFT WITH 43420 OBSERVATIONS OUT OF THE ORIGINAL 64006

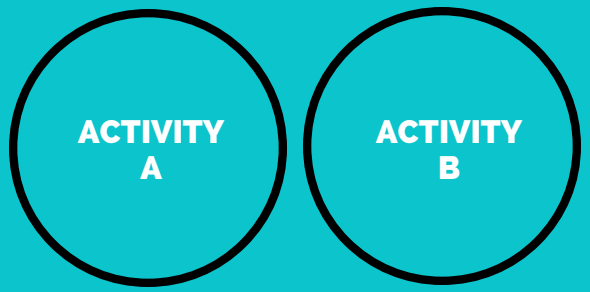
TEAM
RANDOM

PERFORMING CORRECTIONS

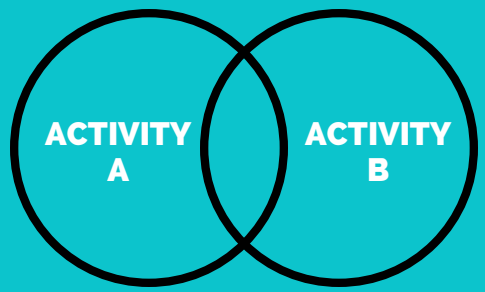
WHAT IF WE TOLD YOU THAT AN INDIVIDUAL SPENT 500 MINUTES SLEEPING IN A 32-HOUR DAY?

THIS HAPPENS BECAUSE OF 3 POSSIBLE SCENARIOS:

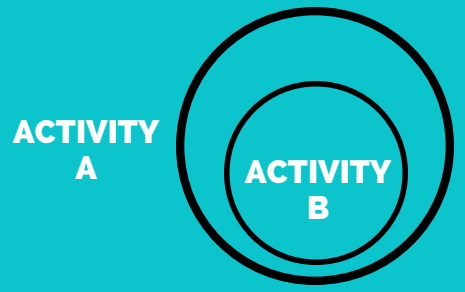
SCENARIO 1



SCENARIO 2



SCENARIO 3

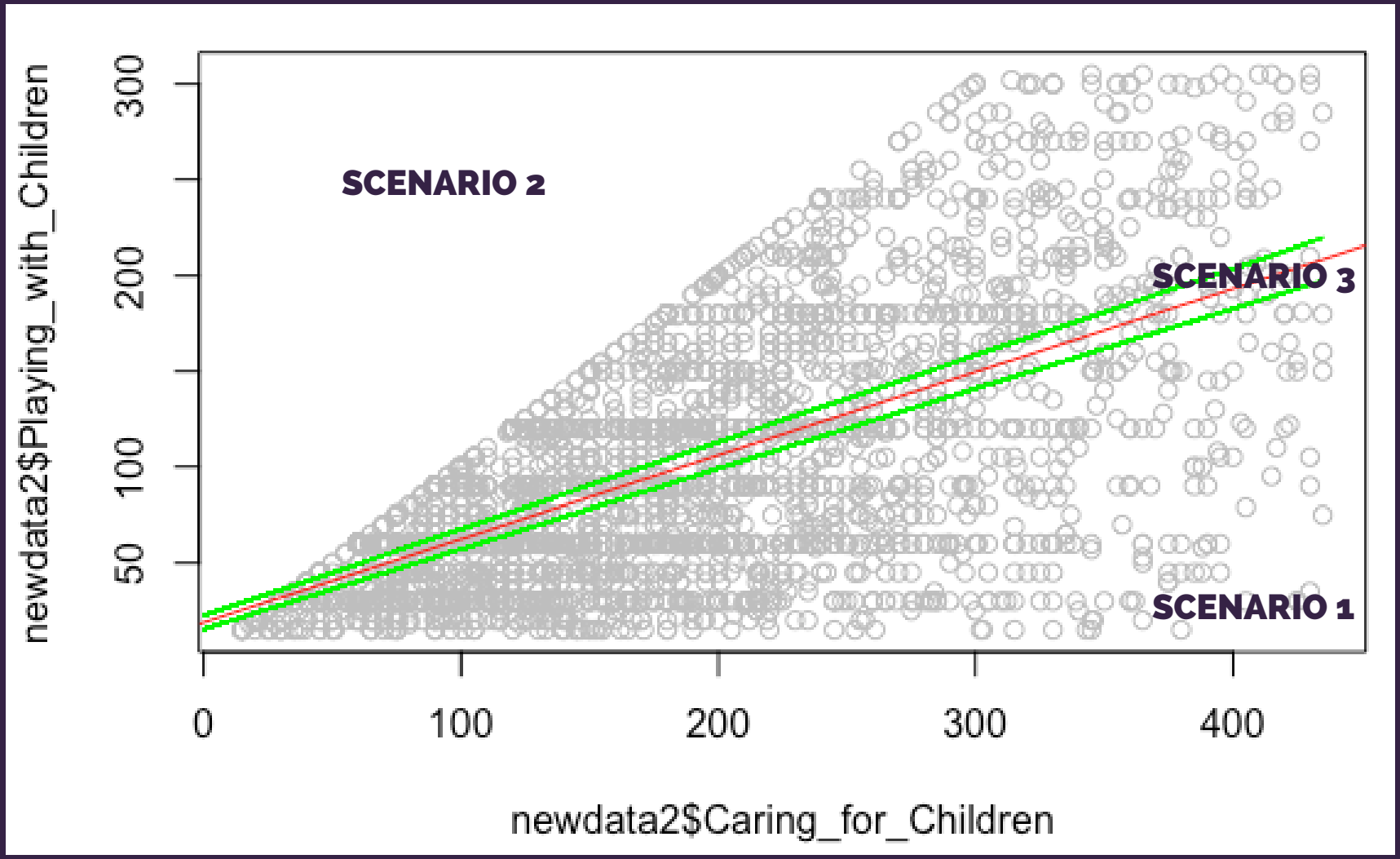


IF NO CORRECTION IS DONE, THERE WILL BE CASES WHERE THE TIME SPENT IS DOUBLE COUNTED. THIS RESULTS IN A DAY LASTING MORE THAN 24 HOURS

CORRECTION CAN ONLY BE DONE ON PAIRS WITH HIGH CORRELATIONS AND EXHIBIT A LINEAR PATTERN.

PLAYING WITH CHILDREN = P + G*CARING WITH CHILDREN **ADJ. R-SQ: 0.404**

INTERPRETATION: PLAYING WITH CHILDREN CAN BE CONSIDERED AS CARING FOR CHILDREN, HOWEVER, NOT ALL ACTIONS OF CARING FOR CHILDREN INCLUDE PLAYING



SCENARIO 1:

SINCE THE STATED TIME IS LESS THAN PREDICTED, IT IS LIKELY THAT THE INDIVIDUAL STATED AS SEPARATE ACTIVITIES

NO CORRECTION NEEDED.

SCENARIO 2:

SINCE THE STATED TIME IS GREATER THAN PREDICTED, IT IS LIKELY THAT THE INDIVIDUAL STATED OVERLAPPING TIME

A - MEAN A.

SCENARIO 3:

SINCE THE STATED TIME IS EQUAL TO PREDICTED, IT IS LIKELY THAT THE INDIVIDUAL STATED A SUBSET TIME

IGNORE A.

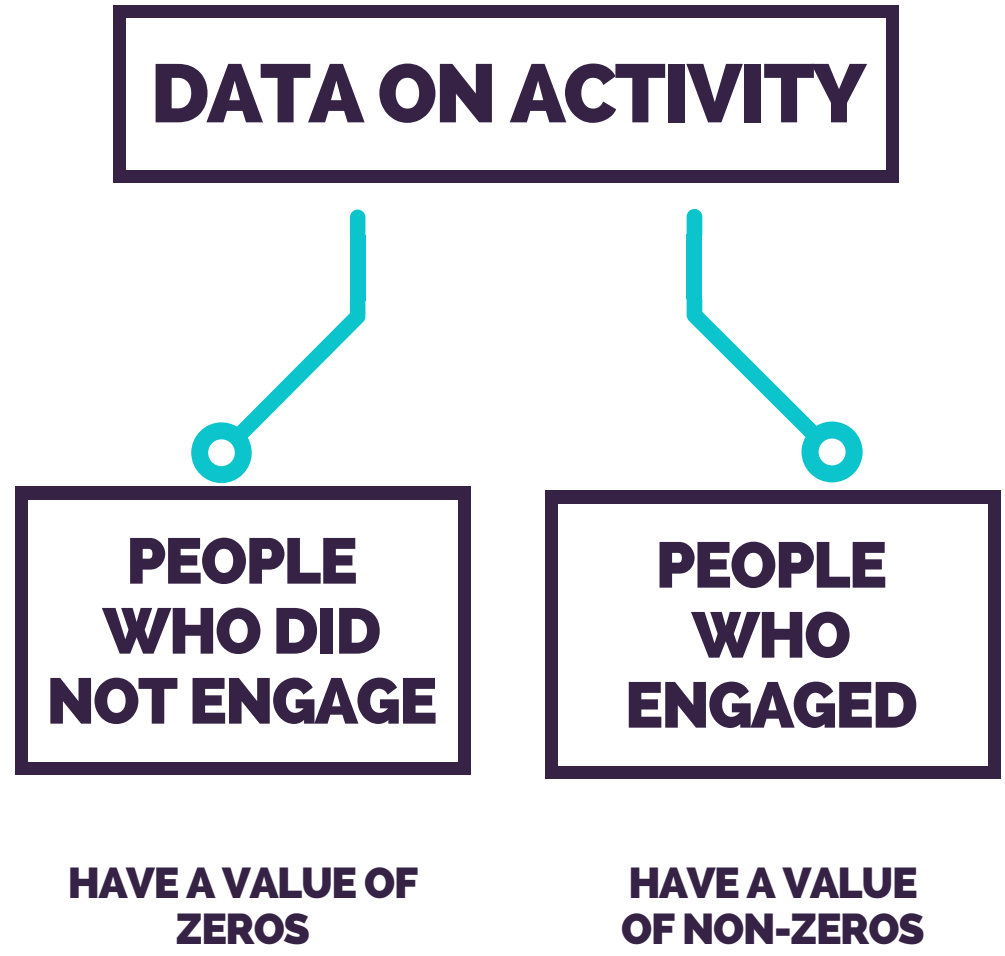
CORRECTED PAIRINGS

WHILE THIS MAY SEEM SMALL, IT AFFECTS THE ACCURACY OF THE MACHINE LEARNING MODELS BY ~20%

1. **HOUSEWORK & FOOD DRINK PREP**
2. **CARING FOR CHILDREN & PLAYING WITH CHILDREN**
3. **TELEVISION & SOCIALIZING RELAXING**
4. **SLEEPING & WEEKLY HOURS WORKED**

WHEN ANALYZING AN ACTIVITY, WE NEED TO BE MINDFUL OF THOSE WHO ENGAGED AND THOSE WHO DID NOT

TEAM RANDOM



IF WE DO NOT ANALYZE THEM SEPARATELY, THESE 2 DISTINCT POPULATIONS WILL MASK THE PATTERNS FROM EACH OTHER.

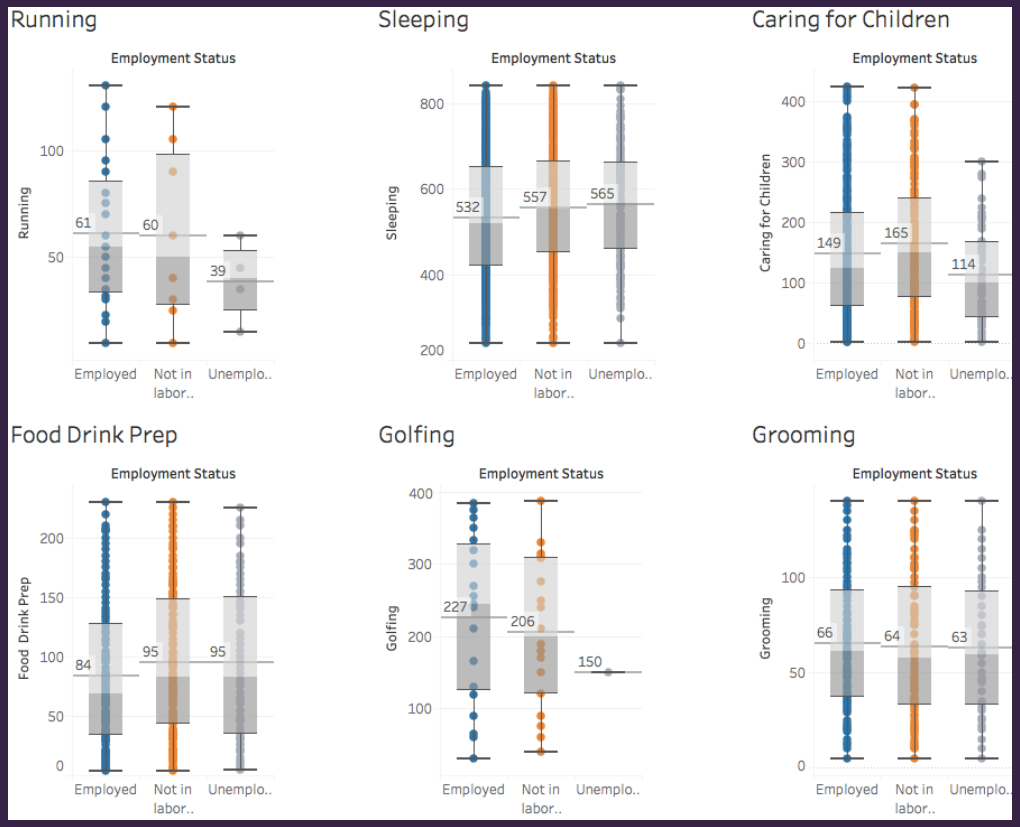
II.

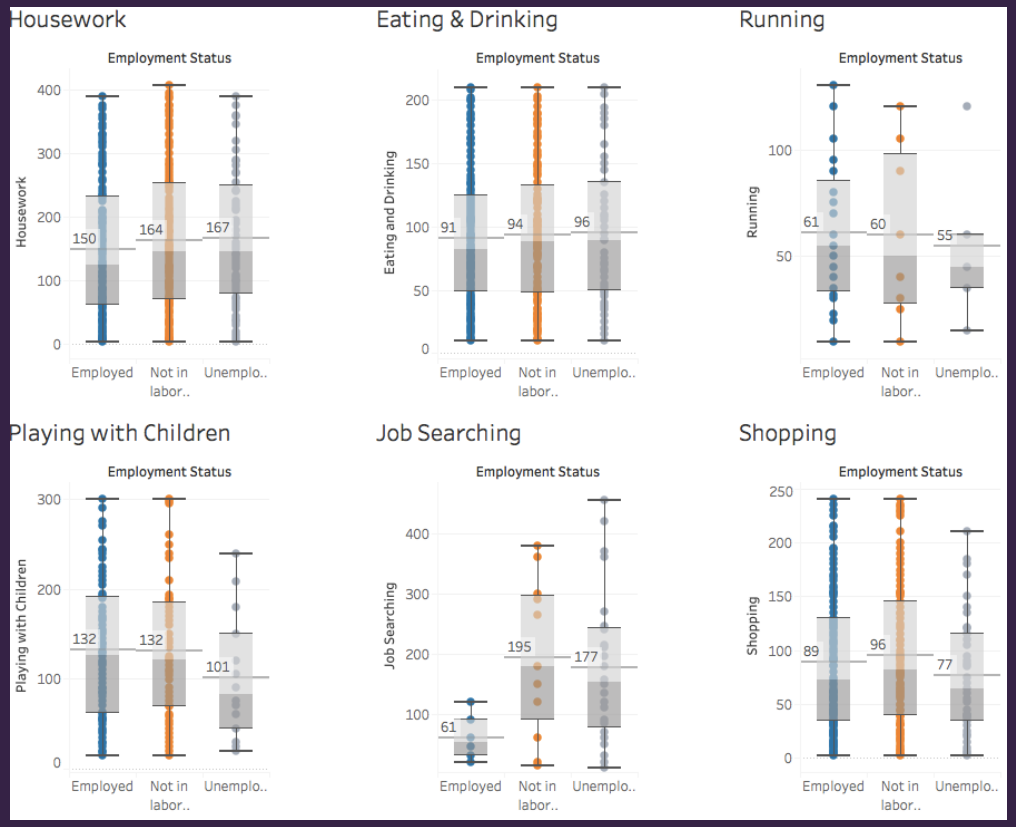
QUESTION 1

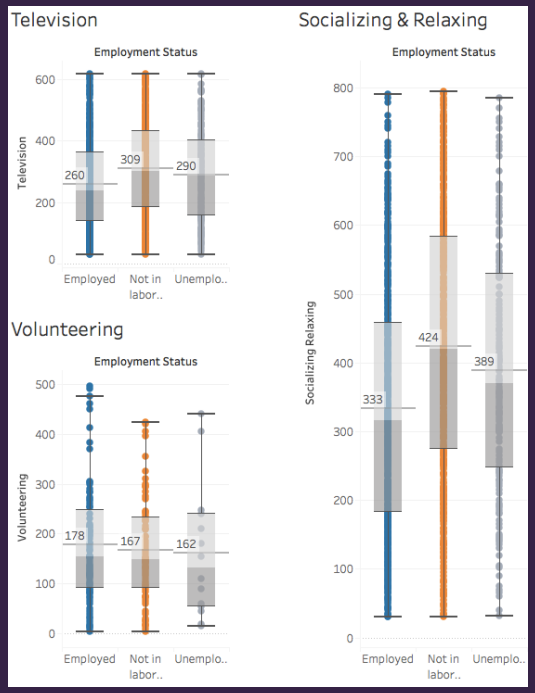
SUMMARY OF TIME SPENDING PATTERN IN 2012 (LIKE TIME SPEND PER ACTIVITY E.G. SOCIALIZING, EATING, WORKING, ETC.)

SINCE THE MAIN QUESTION WE ARE TRYING TO ANSWER DEALS WITH EMPLOYMENT STATUS' INFLUENCE ON TIME SPENDING PATTERNS, WE WILL FOCUS ONLY ON EMPLOYMENT STATUS.

TEAM
RANDOM







2012 BOXPLOTS OF EMPLOYMENT

Through visual inspection, we can see that the variables in 2012:

- The means of each variable appear to be significantly different from each other

To confirm this, we ran ANOVA on time spent in activities against employment status:

- Turns out that the mean of all variables are significantly different from each other EXCEPT for shopping!

INSIGHT: The average shopping time spent is unaffected by employment status (possibly even by recessions).

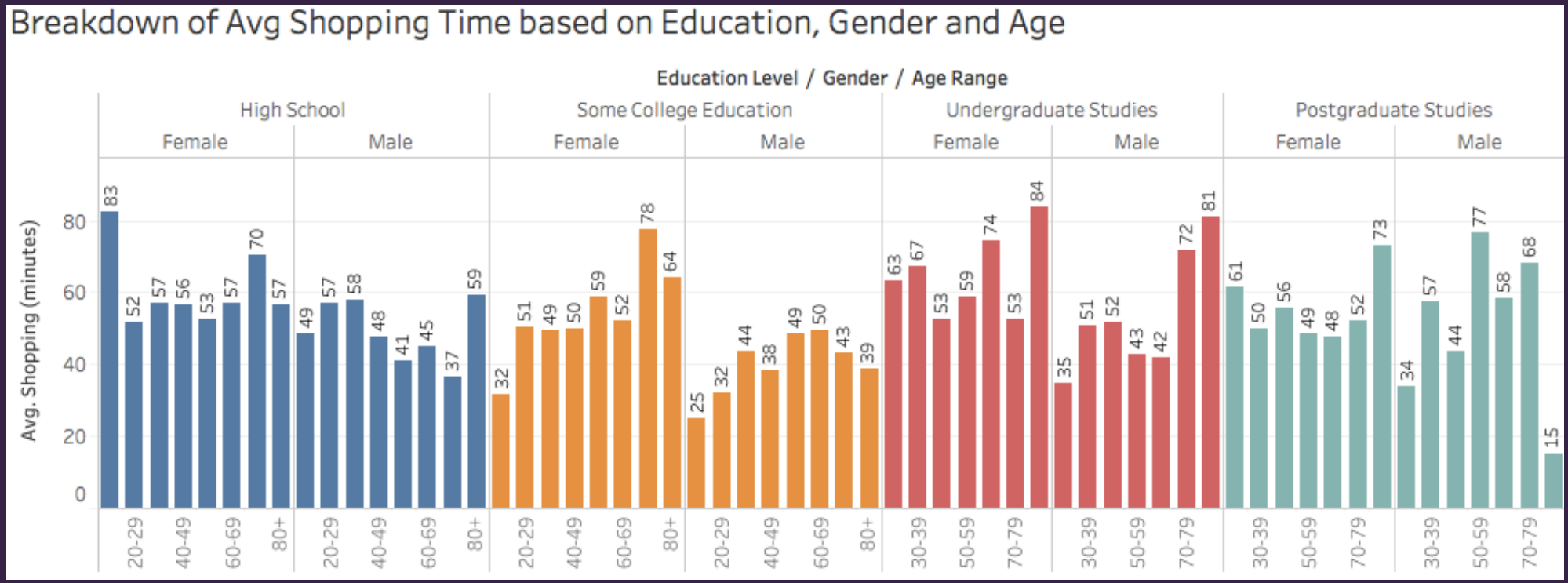
II.

QUESTION 2

HOW DOES THE SPENDING TIME IN QUESTION 1 CHANGES BASED ON AGE, WORKING STATUS, EDUCATION LEVEL ETC.?

ONCE WE REALISED THAT SHOPPING IS UNAFFECTED BY EMPLOYMENT STATUS, WE DECIDED TO FOCUS ONLY ON HOW TIME SPENT SHOPPING CHANGES BASED ON AGE, EDUCATION LEVEL ETC.

TEAM
RANDOM



BREAK-DOWN OF AVERAGE SHOPPING TIME

Through visual inspection, we can see that the mean shopping time changes based on Age, Education and Gender!

To confirm this, we ran ANOVA test.... **Confirms that the means are significantly different from each other!**

INSIGHT: A higher educated male (undergraduate and above) spends an increasing amount of time shopping as he grows older. While a lesser educated male exhibits the converse pattern. On a whole, females tend to spend a longer time shopping in their youth as compared to males, regardless of education level.

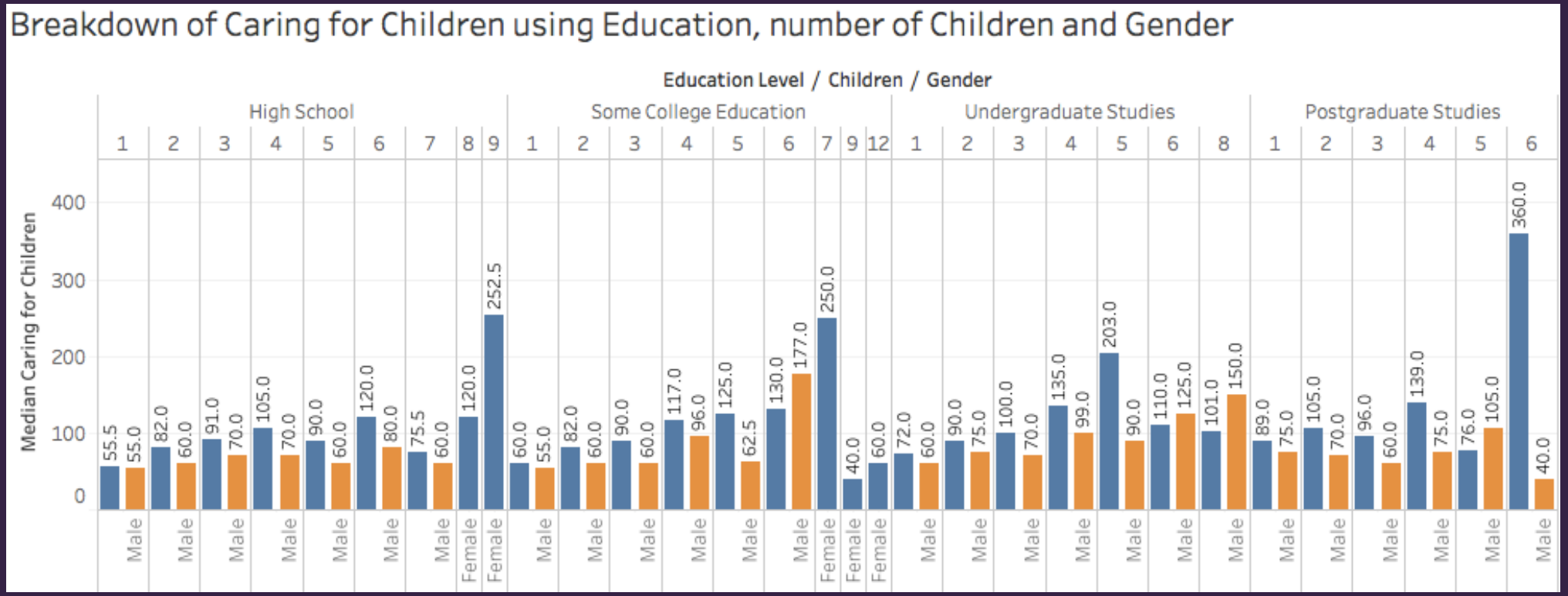
Therefore, as a retailer, a good mix between females and male customers would allow me to reach short and long-term sales goal.

II.

QUESTION 3

HOW TIME SPEND ON BABY CARE IS CHANGING BASED ON EDUCATION, WORKING HOURS, INCOME? (CONSIDER ALL YEARS). OTHER FACTORS CAN ALSO BE CONSIDERED.

TEAM
RANDOM



BREAK-DOWN OF MEDIAN CHILDCARE TIME

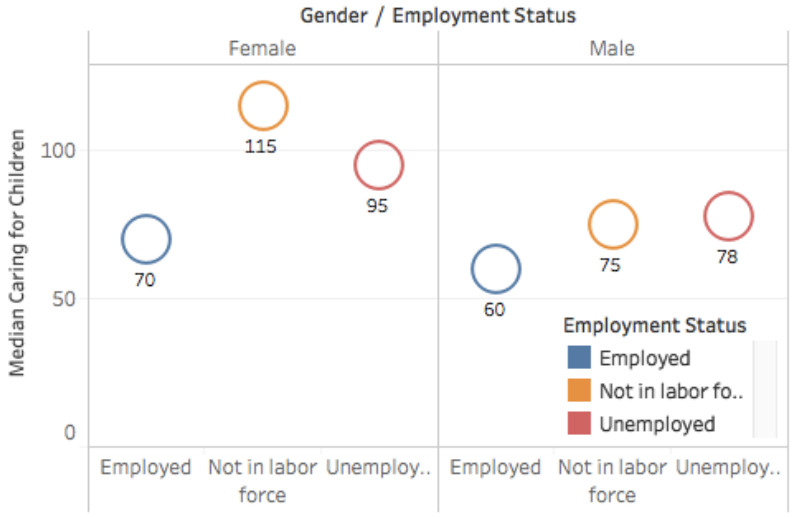
It seems true! That a higher educated individuals have less children than those who are less educated.

Through the results of the ANOVA... **Education, number of Children and Gender are significant contributors to the median childcare time!**

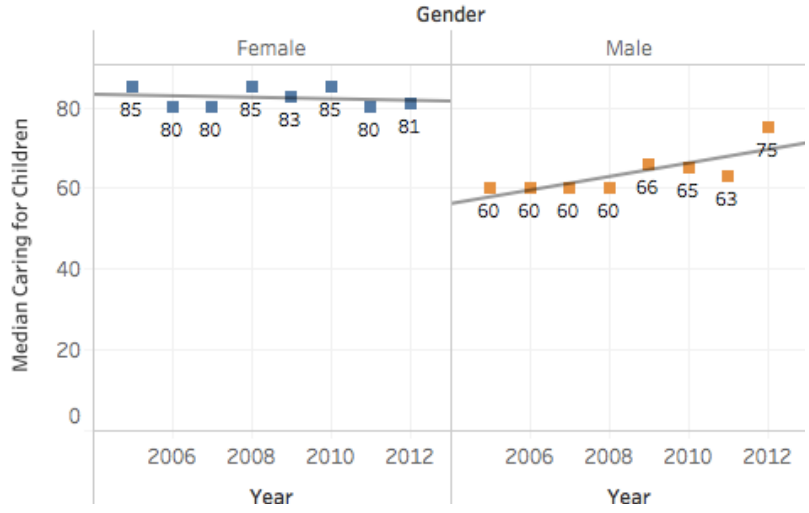
INSIGHT: As the number of children increases, both males and females can be observed to spend more time caring! Looks like males do contribute to looking after kids!

However, there is a noticeable threshold where once the number of children is exceeded (6-8), males just give up and spend 0 time looking after kids! Therefore, don't have more than 6 kids and both parents will contribute equally!

Breakdown of Median Childcare based on Gender and Employment



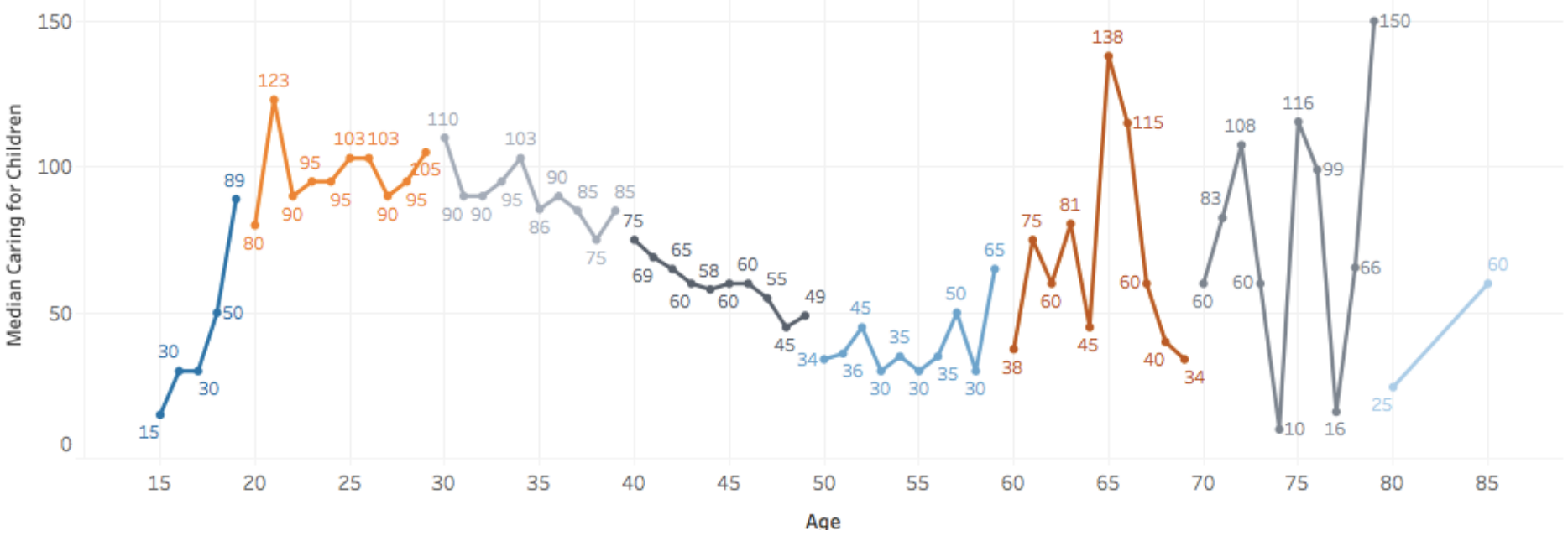
Yearly Trend between Gender and Caring for Children



Regardless of Gender, employment reduces time spent on childcare.

While females have been strangely constant, males have been improving their childcare skills

Age



The main child bearing age is 20-39. As the age of the child increases, the time spent goes on decreasing.The second spike indicates grandparents spending time with grandchildren.

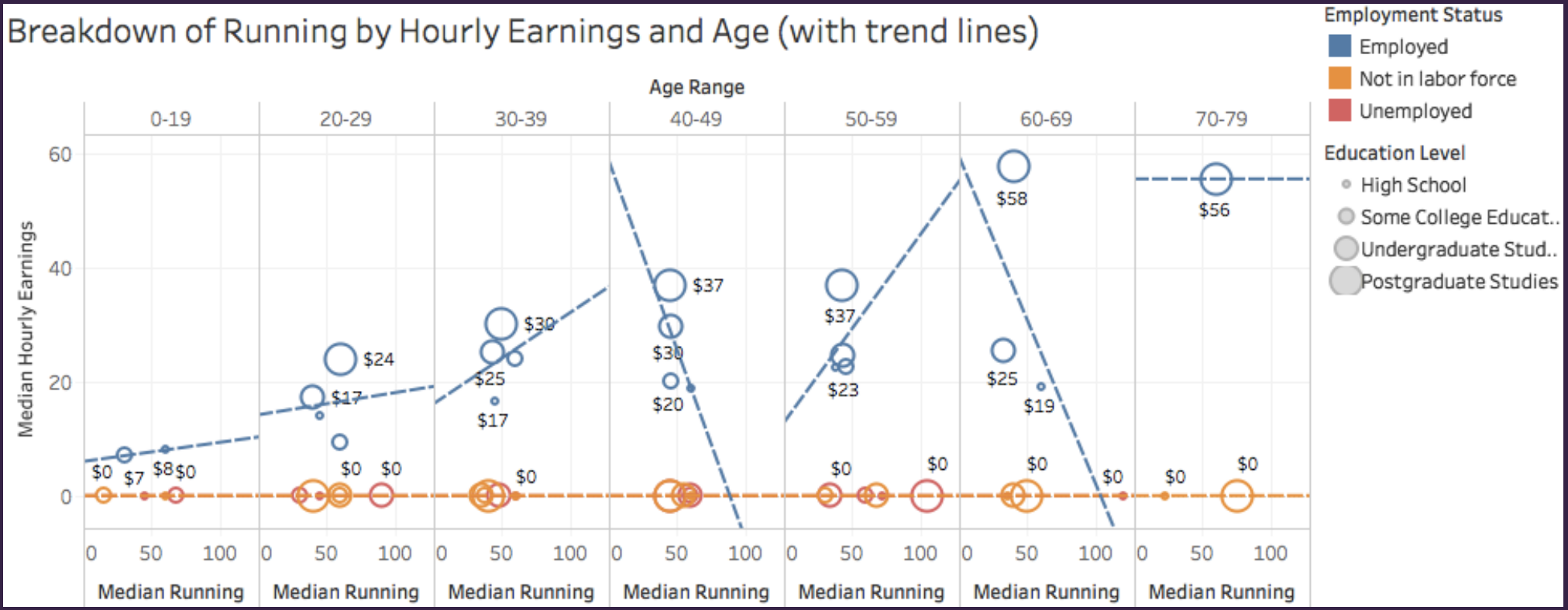
II.

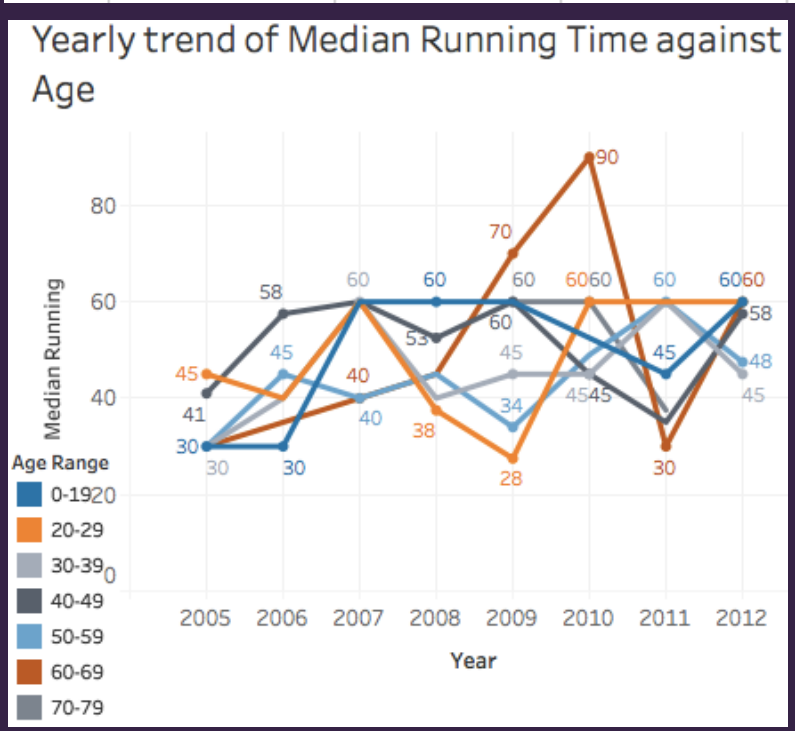
QUESTION 4

HOW IS LEISURE TIME CHANGING BASED ON INCOME AND IS THERE A DIFFERENCE BETWEEN GENERATIONS LEISURE SPENDING TIME?

THE LEISURE ACTIVITY IN FOCUS IS RUNNING

TEAM
RANDOM





TRENDS OF MEDIAN RUNNING

From the graph on the left, it isn't true that aged people exercise any lesser than a youth (0-19 years old).

From the graph on the top, it would seem that in 4 out of the 7 possible age range, having an increased median running time is associated with an increase in median hourly earnings.

INSIGHT: Therefore, as a sports goods manager, I would design and sell higher priced running shoes to the 4 age groups, using the marketing story of Running = Success.

II.

QUESTION 5

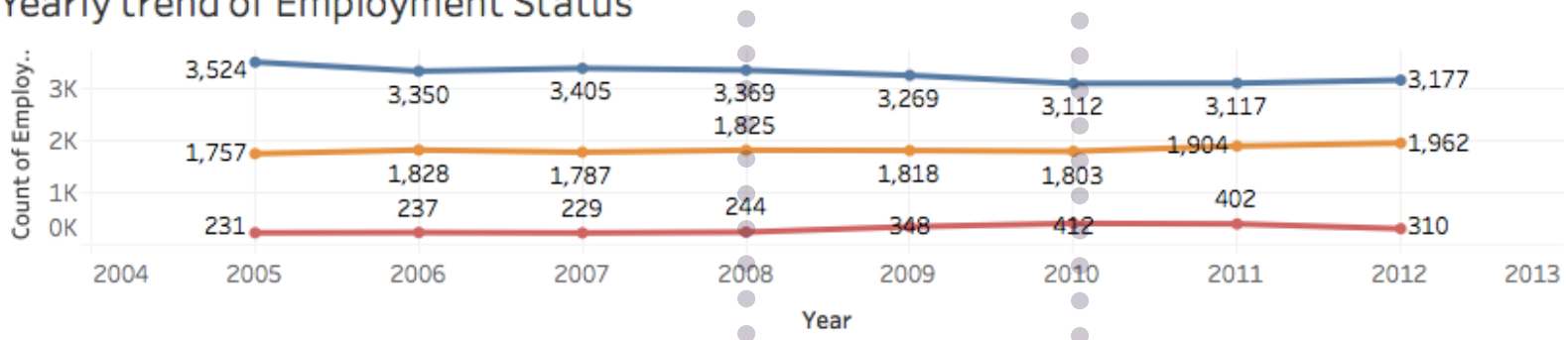
IS THERE ANY CHANGE IN THE PATTERN WHEN THE GREAT RECESSION HAPPENED? IF YES, THEN HAS THE PATTERN CONTINUED AFTER THE RECESSION?

RECESSION IS DEFINED HERE AS A DROP IN EMPLOYMENT AND AN INCREASE IN UNEMPLOYMENT

TEAM
RANDOM

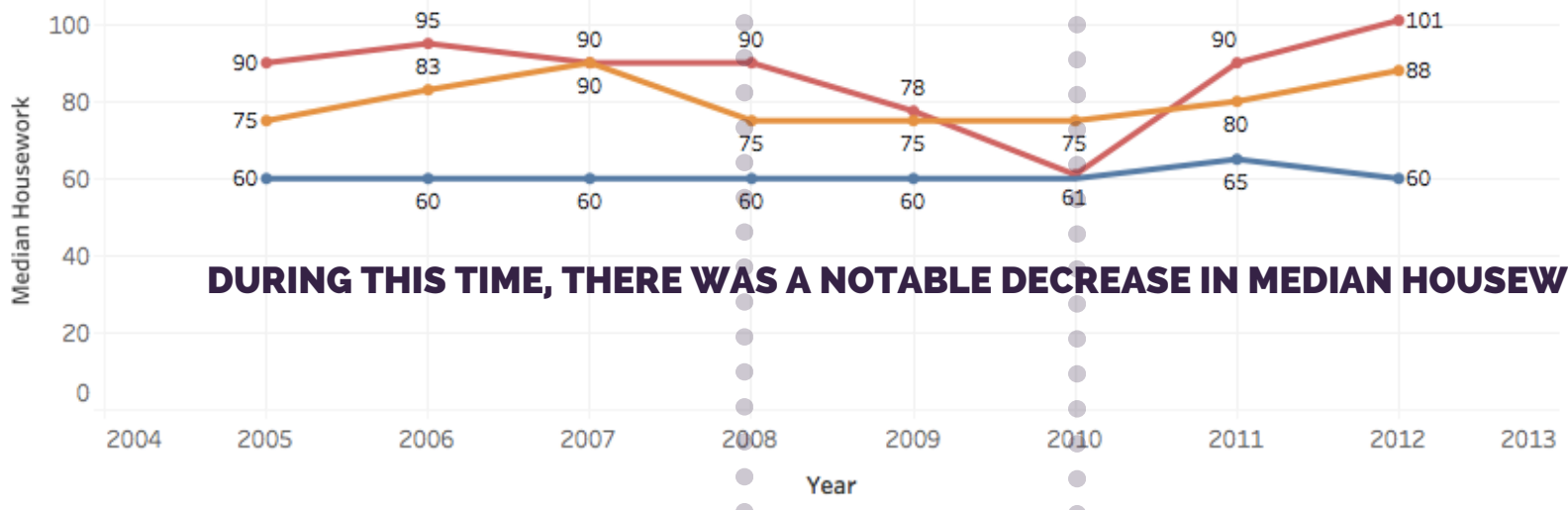
PATTERNS DURING RECESSION

Yearly trend of Employment Status



DECREASE IN EMPLOYMENT LEVELS DURING RECESSION

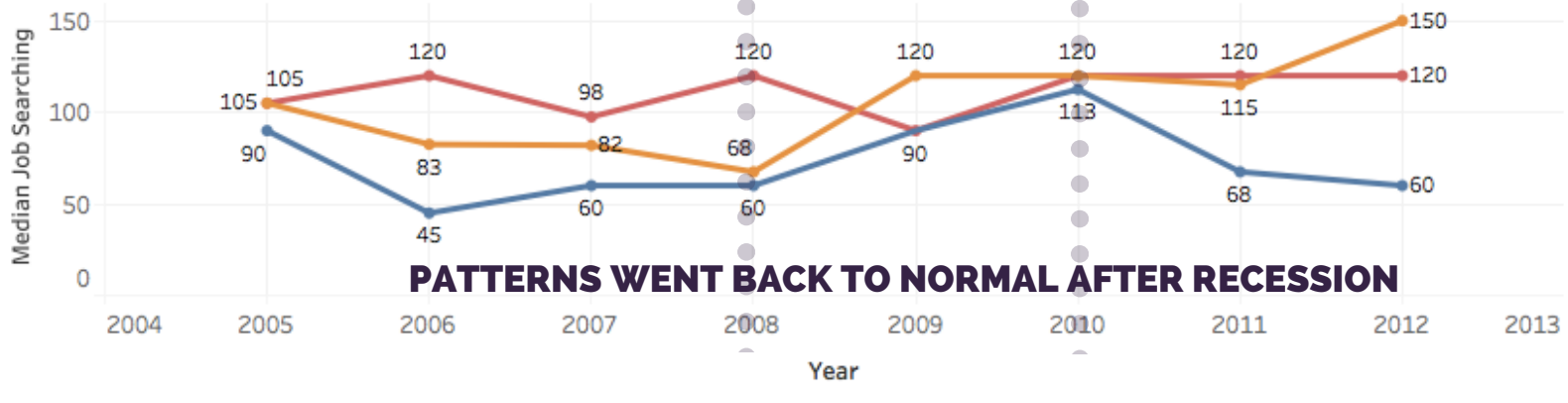
Housework



DURING THIS TIME, THERE WAS A NOTABLE DECREASE IN MEDIAN HOUSEWORK

Job Searching

JOB-SEARCH DRASTICALLY INCREASED AMONG THE EMPLOYED



PATTERNS WENT BACK TO NORMAL AFTER RECESSION

II.

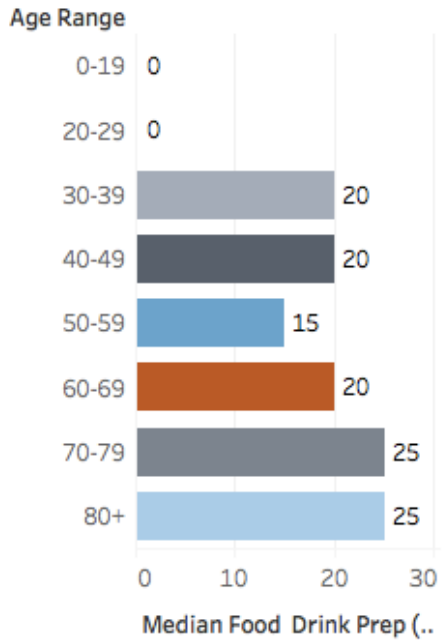
QUESTION 6

BASED ON AGE WHAT IS AN INDIVIDUAL'S PRIMARY ACTIVITY?

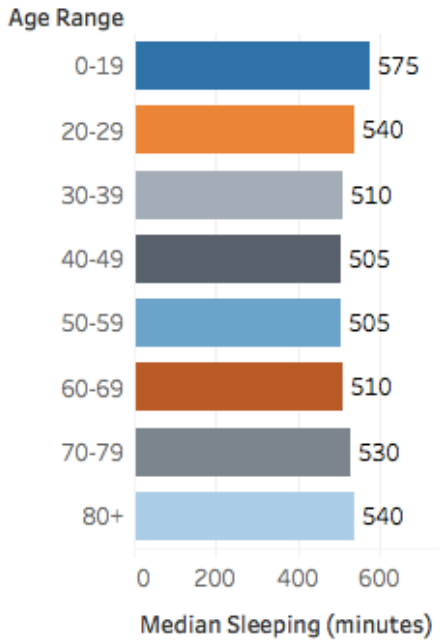
PRIMARY ACTIVITY IS DEFINED AS THE HIGHEST NUMBER OF MINUTES SPENT WHEN COMPARED AGAINST OTHER ACTIVITIES.

TEAM RANDOM

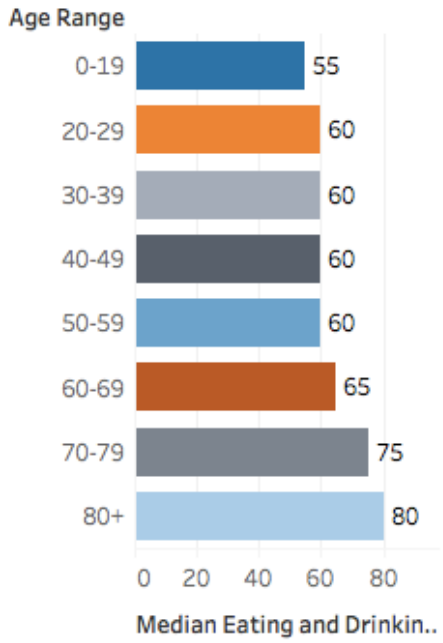
Food Drink Prep



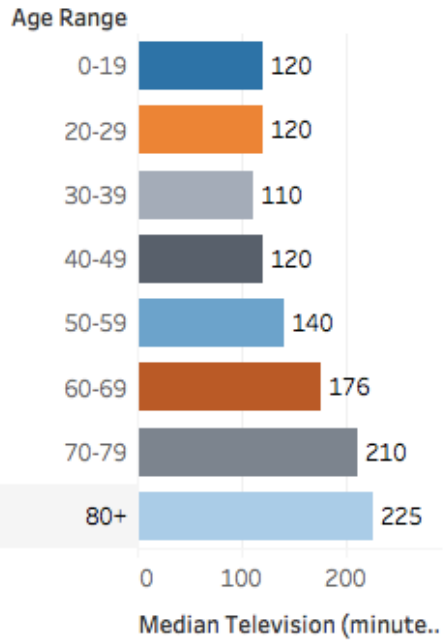
Sleeping



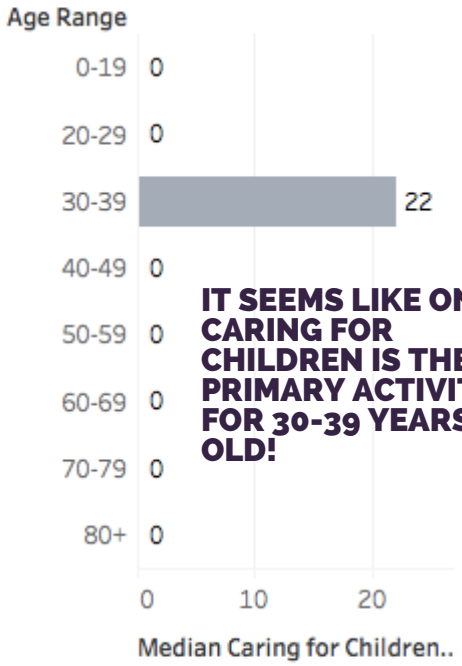
Eating and Drinking



Television

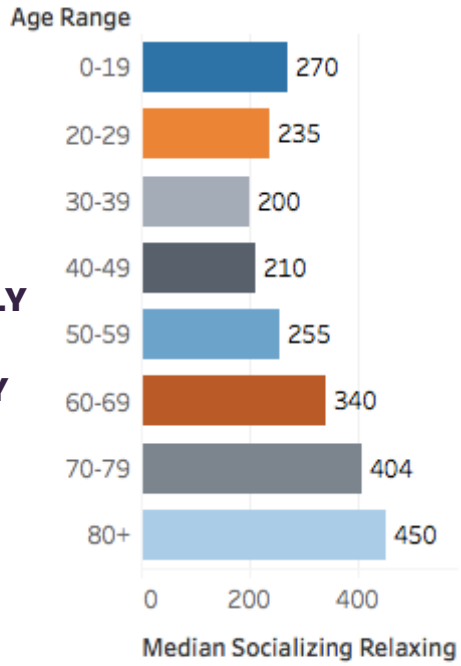


Caring for Children

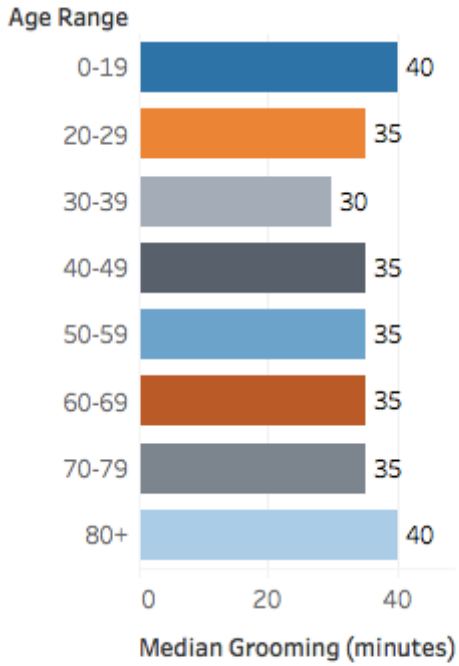


IT SEEMS LIKE ONLY CARING FOR CHILDREN IS THE PRIMARY ACTIVITY FOR 30-39 YEARS OLD!

Socializing & Relaxing



Grooming



- **REGARDLESS OF AGE, THE PRIMARY ACTIVITY HAS ALWAYS BEEN SLEEPING.**
- **FOLLOWED BY SOCIALIZING AND RELAXING.**

INSIGHT:

It is interesting to note that regardless of age, the priorities of activities remain the same.

Therefore, using activities as a predictor of age may not be such a good idea!

As a business man, my marketing strategy for the above activities should not be based using age solely.

II.

QUESTION 7 & 8

WHICH IS THE MOST SIGNIFICANT VARIABLE AFFECTING EMPLOYMENT AS WELL AS UNEMPLOYMENT?

THE PRIMARY FOCUS WILL BE ON USING TIME SPENT ON ACTIVITIES AND ITS EFFECTS ON EMPLOYMENT STATUS

TEAM
RANDOM

UNCORRECTED VARIABLES >>>>>>>>

1. SOCIALIZING_RELAXING
2. SLEEPING
3. FOOD AND DRINK PREP
4. EATING_AND_DRINKING
5. CARING_FOR_CHILDREN
6. HOUSEWORK
7. GROOMING
8. JOB_SEARCHING
9. SHOPPING
10. TELEVISION
11. VOLUNTEERING
12. RUNNING
13. PLAYING_WITH_CHILDREN
14. GOLFING

>>>>>>>> CORRECTED VARIABLES

1. JOB_SEARCHING
2. CORRECTED_PLAYING
3. RUNNING
4. GOLFING
5. VOLUNTEERING
6. CARING_FOR_CHILDREN
7. CORRECTED_FOODDRINKPREP
8. HOUSEWORK
9. SHOPPING
10. CORRECTED_SOCIALIZING
11. CORRECTED_SLEEPING
12. EATING_AND_DRINKING
13. CORRECTED_TV
14. GROOMING

JUST BY USING THE CORRECTED VARIABLES, WHICH ACCOUNTS FOR TIME OVERLAP

FOR THE SAME MODEL, WE CAN EXPERIENCE

~20%

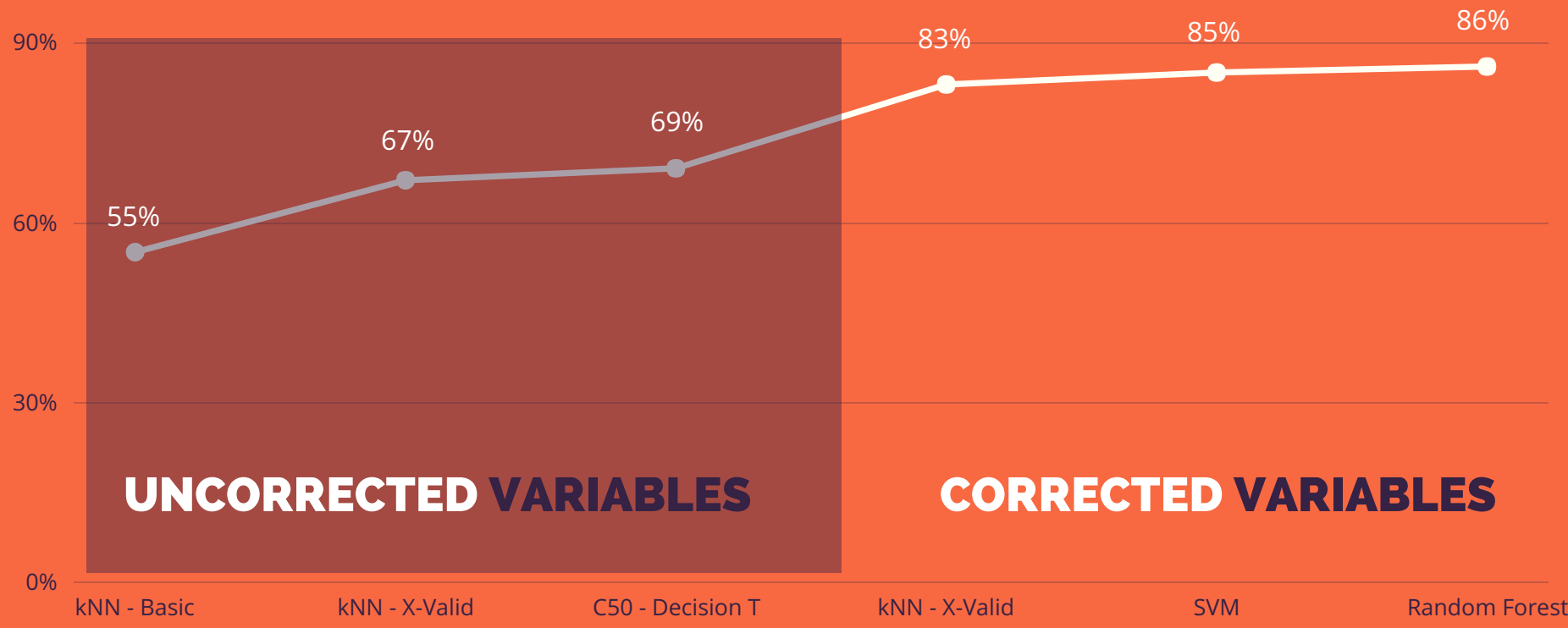
IMPROVEMENT

III.

MODEL RESULTS

The results reported are based on testing the model on the training dataset (after partitioning).

A total of 4 different types of models with 6 different combinations (i.e. without pre-processing, without cross-validation etc.)



MACHINE LEARNING MODEL SELECTION

WE CHOSE SVM OVER RANDOM FOREST BECAUSE

SVM performs better in terms of misclassifications (by almost 50% in false positives).

It is imperative that we do not misclassify the “employed” or “not in labor force” as “unemployed”.

THIS ALLOWS OUR MODEL TO FIND TIME-SPENDING PATTERNS DURING RECESSION MORE EFFECTIVELY.

THE UNIVERSITY OF TEXAS AT DALLAS

THANK YOU

HELPING THE WORLD, ONE DATASET AT A TIME

TEAM RANDOM