**BRFSS2015**

Q1. How many people have any kind of health care coverage?

Sol: There are **407556** people have any kind of health care coverage.

Q2. What is the average "Number of Days Mental Health Not Good" for those in Pennsylvania who have numeric data?

Sol: **3.32 Average Days**

Q3. Compare only those who have and have not had some form of arthritis, rheumatoid arthritis, gout, etc. in the format of

mean\_weight sd\_weight

183.04 xx.xx

xxx.xx xx.xx

Sol:

HAVARTH3 mean\_weight sd\_weight

1 1 183.04 **49.81**

2 2 **176.08 46.31**

Q4. Remove outliers from minutes of total physical activity per week using 0.997 and 0.003 as criteria. What percentage of observations remain?

Sol: [1] 99.7%

Q5. Group by marital status and calculate the mean, standard deviation, minimum, and maximum of total exercise, to two decimals.

Sol: A tibble: 6 x 5

MARITAL total\_ex\_mean total\_ex\_std total\_ex\_min total\_ex\_max

1 1 486. 727. 0 47640

2 2 491. 835. 0 54000

3 3 592. 854. 0 27360

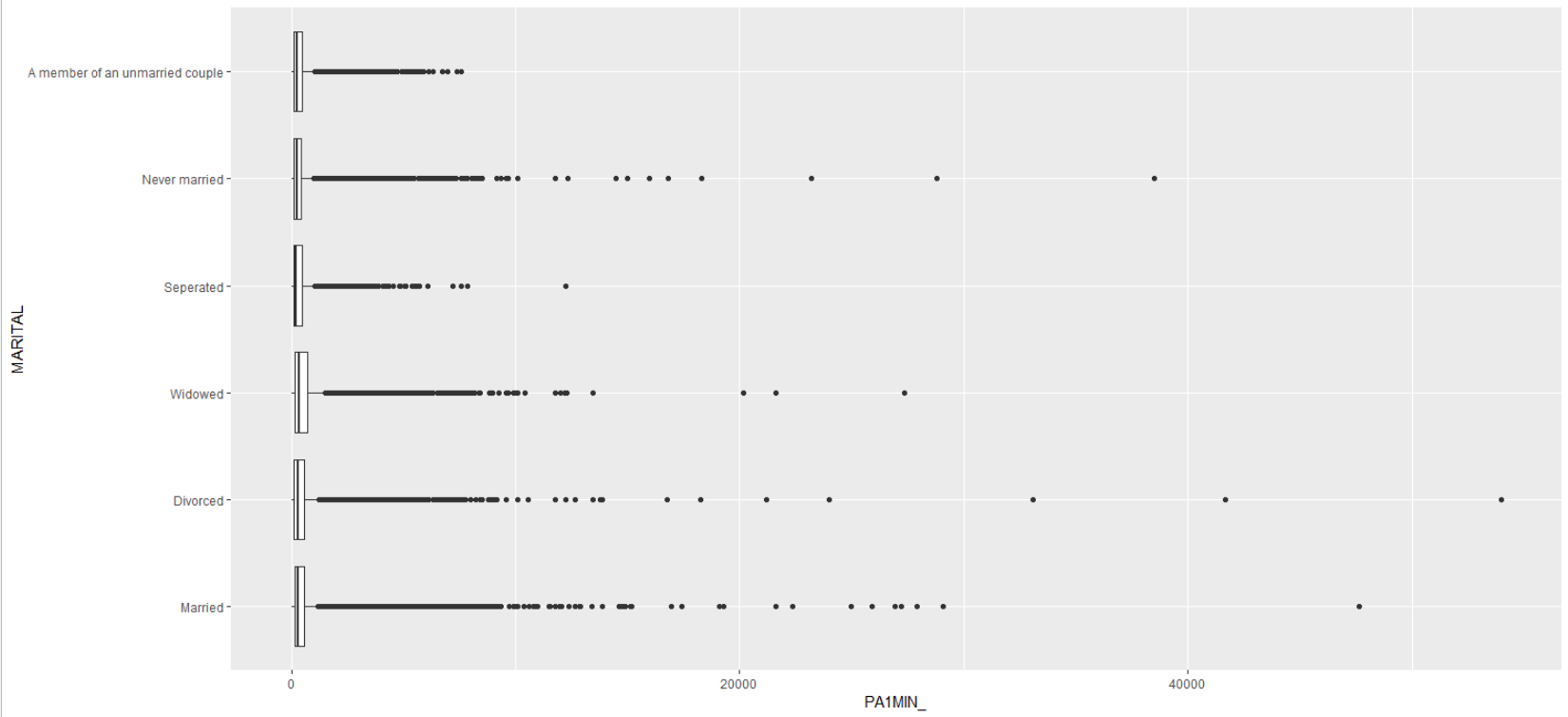
4 4 419. 627. 0 12240

5 5 407. 662. 0 38520

6 6 432. 602. 0 7560

Q6. Create a boxplot for total exercise by marital status.

Sol:



Q7. Run a regression predicting exercise by marital status. Assign the model summary to Q7

Sol: Call:

lm(formula = PA1MIN\_ ~ MARITAL, data = convert\_fctr\_data)

Residuals:

Min 1Q Median 3Q Max

-592 -351 -206 69 53509

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 486.235 1.853 262.397 < 2e-16 \*\*\*

MARITAL2 4.885 4.298 1.137 0.256

MARITAL3 106.223 4.535 23.424 < 2e-16 \*\*\*

MARITAL4 -66.890 10.576 -6.325 2.54e-10 \*\*\*

MARITAL5 -79.006 3.966 -19.920 < 2e-16 \*\*\*

MARITAL6 -54.655 8.270 -6.608 3.89e-11 \*\*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 742.5 on 287791 degrees of freedom

Multiple R-squared: 0.00433, Adjusted R-squared: 0.004313

F-statistic: 250.3 on 5 and 287791 DF, p-value: < 2.2e-16

Q8. Run an ANOVA comparing exercise across marital status, and assign the TukeyHSD post-hoc test to Q8.

Sol:

**ANOVA**

aov(formula = PA1MIN\_ ~ MARITAL, data = convert\_fctr\_data)

Terms:

MARITAL Residuals

Sum of Squares 689973238 158656937197

Deg. of Freedom 5 287791

Residual standard error: 742.4905

Estimated effects may be unbalanced

**TukeyHSD:**

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = PA1MIN\_ ~ MARITAL, data = convert\_fctr\_data)

$MARITAL

diff lwr upr p adj

2-1 4.885433 -7.3617145 17.13258 0.8661197

3-1 106.222649 93.2996888 119.14561 0.0000000

4-1 -66.889901 -97.0280353 -36.75177 0.0000000

5-1 -79.005863 -90.3083280 -67.70340 0.0000000

6-1 -54.654509 -78.2225729 -31.08645 0.0000000

3-2 101.337216 85.1747574 117.49967 0.0000000

4-2 -71.775334 -103.4380771 -40.11259 0.0000000

5-2 -83.891296 -98.7898825 -68.99271 0.0000000

6-2 -59.539942 -85.0286807 -34.05120 0.0000000

4-3 -173.112549 -205.0427793 -141.18232 0.0000000

5-3 -185.228511 -200.6874296 -169.76959 0.0000000

6-3 -160.877158 -186.6974222 -135.05689 0.0000000

5-4 -12.115962 -43.4254239 19.19350 0.8804463

6-4 12.235392 -25.2878052 49.75859 0.9390884

6-5 24.351354 -0.6971769 49.39988 0.0622828

Q9: Run a regression as in Q7, but add total fruits consumed per day. Based on the R-squared and AIC, what is the better model? Assign the better AIC value to Q9.

Sol: Call:

lm(formula = PA1MIN\_ ~ MARITAL + FRUITSUM, data = Convert\_new)

Residuals:

Min 1Q Median 3Q Max

-3559 -352 -208 65 53477

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 474.25770 2.78952 170.01 <2e-16 \*\*\*

MARITAL -12.00872 0.86740 -13.84 <2e-16 \*\*\*

FRUITSUM 0.24247 0.01011 23.97 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 738.6 on 281059 degrees of freedom

(6735 observations deleted due to missingness)

Multiple R-squared: 0.002781, Adjusted R-squared: 0.002773

F-statistic: 391.8 on 2 and 281059 DF, p-value: < 2.2e-16

> modelBest\_possible <- ols\_step\_all\_possible(data\_mod2)

> modelBest\_possible

Index N Predictors R-Square Adj. R-Square Mallow's Cp

2 1 1 FRUITSUM 0.0021004880 0.002096937 192.6702

1 2 1 MARITAL 0.0007798379 0.000776366 4067.1143

3 3 2 MARITAL FRUITSUM 0.0027805488 0.002773453 3.0000

> modelBest\_subset <- ols\_step\_best\_subset(data\_mod2)

> modelBest\_subset

Best Subsets Regression

-------------------------------

Model Index Predictors

-------------------------------

1 FRUITSUM

2 MARITAL FRUITSUM

-------------------------------

Subsets Regression Summary

-------------------------------------------------------------------------------------------------------------------------------------------------------------

Adj. Pred

Model R-Square R-Square R-Square C(p) AIC SBIC SBC MSEP FPE HSP APC

-------------------------------------------------------------------------------------------------------------------------------------------------------------

1 0.0021 0.0021 0.0021 192.6702 4510513.5853 3712894.1786 4510545.2243 153435814859.6937 545918.3621 1.9423 0.9979

2 0.0028 0.0028 0.0027 3.0000 4510323.9784 3712704.5744 4510366.1637 153331795088.8426 545550.2050 1.9410 0.9972

-------------------------------------------------------------------------------------------------------------------------------------------------------------

AIC: Akaike Information Criteria

SBIC: Sawa's Bayesian Information Criteria

SBC: Schwarz Bayesian Criteria

MSEP: Estimated error of prediction, assuming multivariate normality

FPE: Final Prediction Error

HSP: Hocking's Sp

APC: Amemiya Prediction Criteria

> Q9 <- AIC(data\_mod2, k = 2)

> Q9

**[1] 4510324**

**For the final section, you will choose four variables to explore that we previously have not.**

**Complete the following;**

● Q10: Address the values of any variables. For instance, is “none” equal to a value other

than 0? Are there extra decimals implied?

Sol: Q10

# A table: 393,866 x 4

RFSEAT3 EMPLOY1 CHLDCNT RFHLTH

<dbl> <dbl> <dbl> <dbl>

1 1 8 1 2

2 2 3 1 1

3 1 8 2 2

4 1 8 1 2

5 1 2 1 1

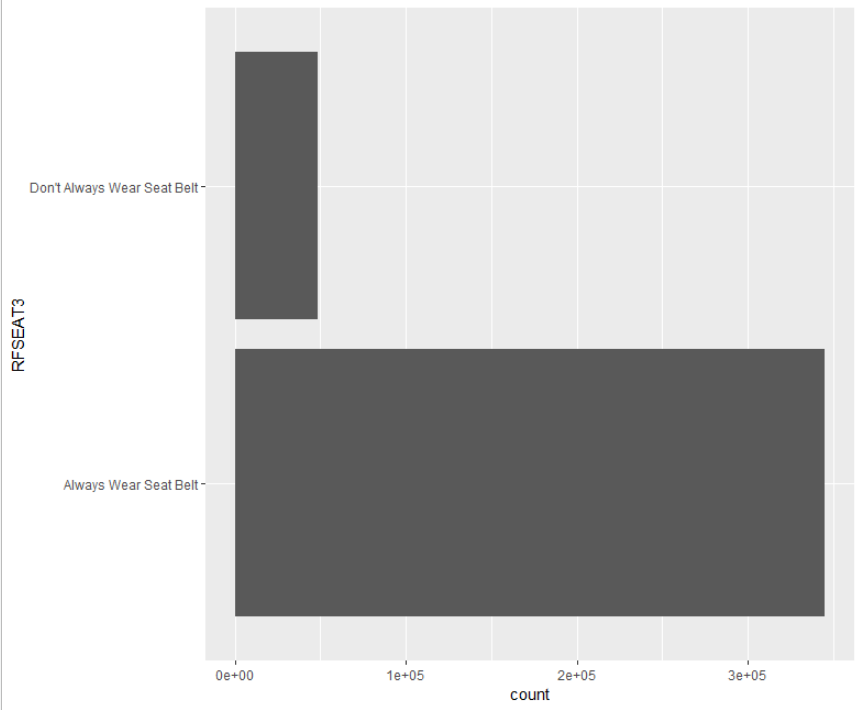
6 1 7 1 1

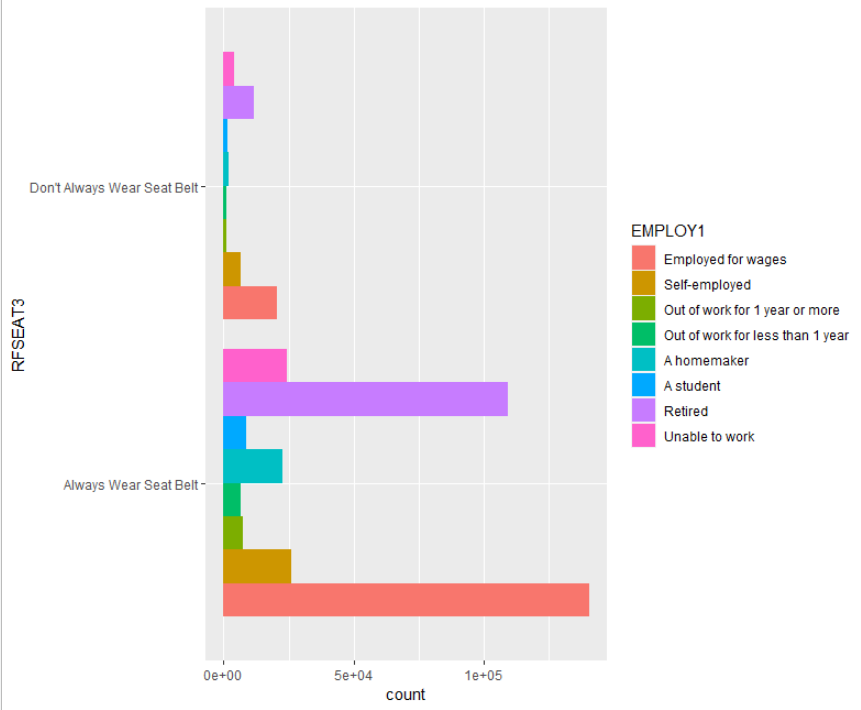
7 1 3 1 2

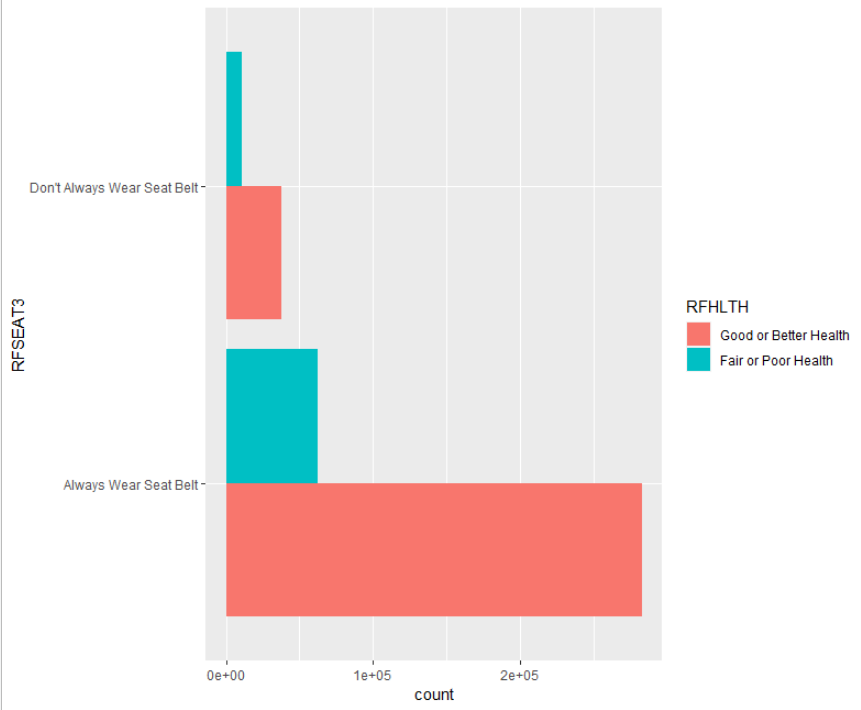
8 1 5 1 2

● Q12: Complete exploratory analyses doing appropriate visualizations with ggplot2.

Sol:







*(Few visualizations are here for more details please refer .R file)*

● Q13: Run basic descriptive statistics

Sol:

vars n mean sd median trimmed mad min max range skew kurtosis se

RFSEAT3\* 1 393866 1.12 0.33 1 1.03 0.00 1 2 1 2.29 3.24 0

EMPLOY1\* 2 393866 3.91 2.85 3 3.80 2.97 1 8 7 0.16 -1.79 0

CHLDCNT\* 3 393866 1.50 0.99 1 1.26 0.00 1 6 5 2.17 4.41 0

RFHLTH\* 4 393866 1.19 0.39 1 1.11 0.00 1 2 1 1.62 0.62 0

● Q14: Finally, run an appropriate regression predicting one of those variables. Identify

the best model.

Sol:

By glm command logistic regression is performed to select best model:

AIC values :

[1] 327686.6 374707.1 375113.9 377464.8 328560.0

Model 1 is best among all the models:

**Rfhlthbin ~ CHLDCNT + RFSEAT3 + EMPLOY1**

AIC value of model1 is 327686.6 which is minimum between all other models.