Computational Assigment #1

Brandon Moretz

### 1.) Given the variables in this dataset, which variables can be considered explanatory (X) and which considered response (Y)? Can any variables take on both roles? What is the population of interest for this problem (yes – this is a trick question!)?

Explanatory variables:

* High School
* College
* Smokers
* Obese
* Heavy Drink

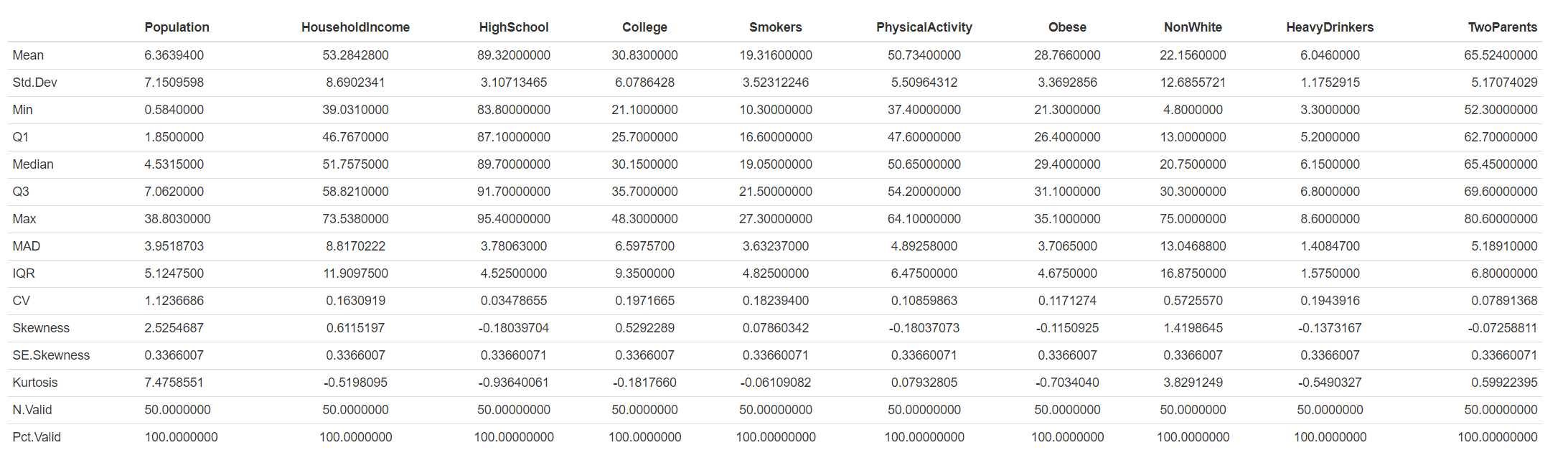
Response variables:

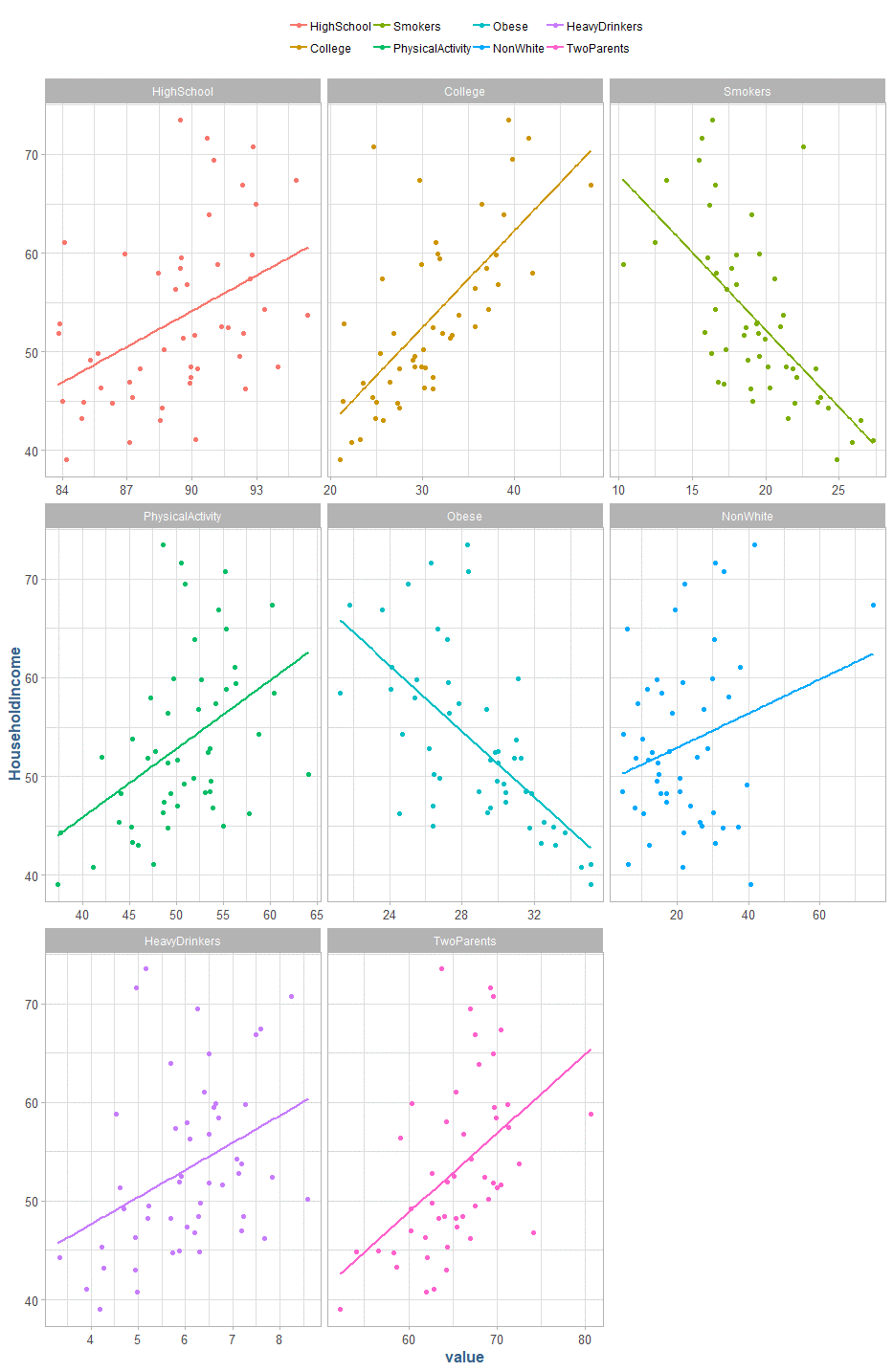
Both:

* Two Parents
* Heavy Drinkers
* Household Income
* Physical Activity

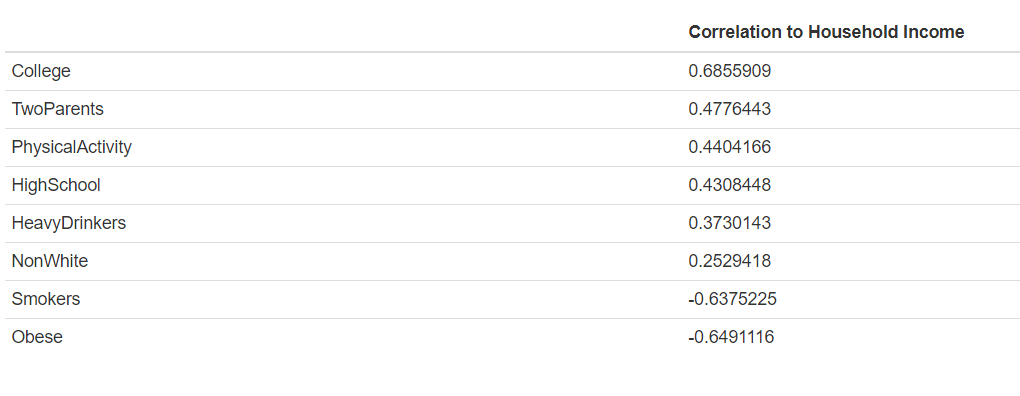
The population of interest for this dataset is the US population.

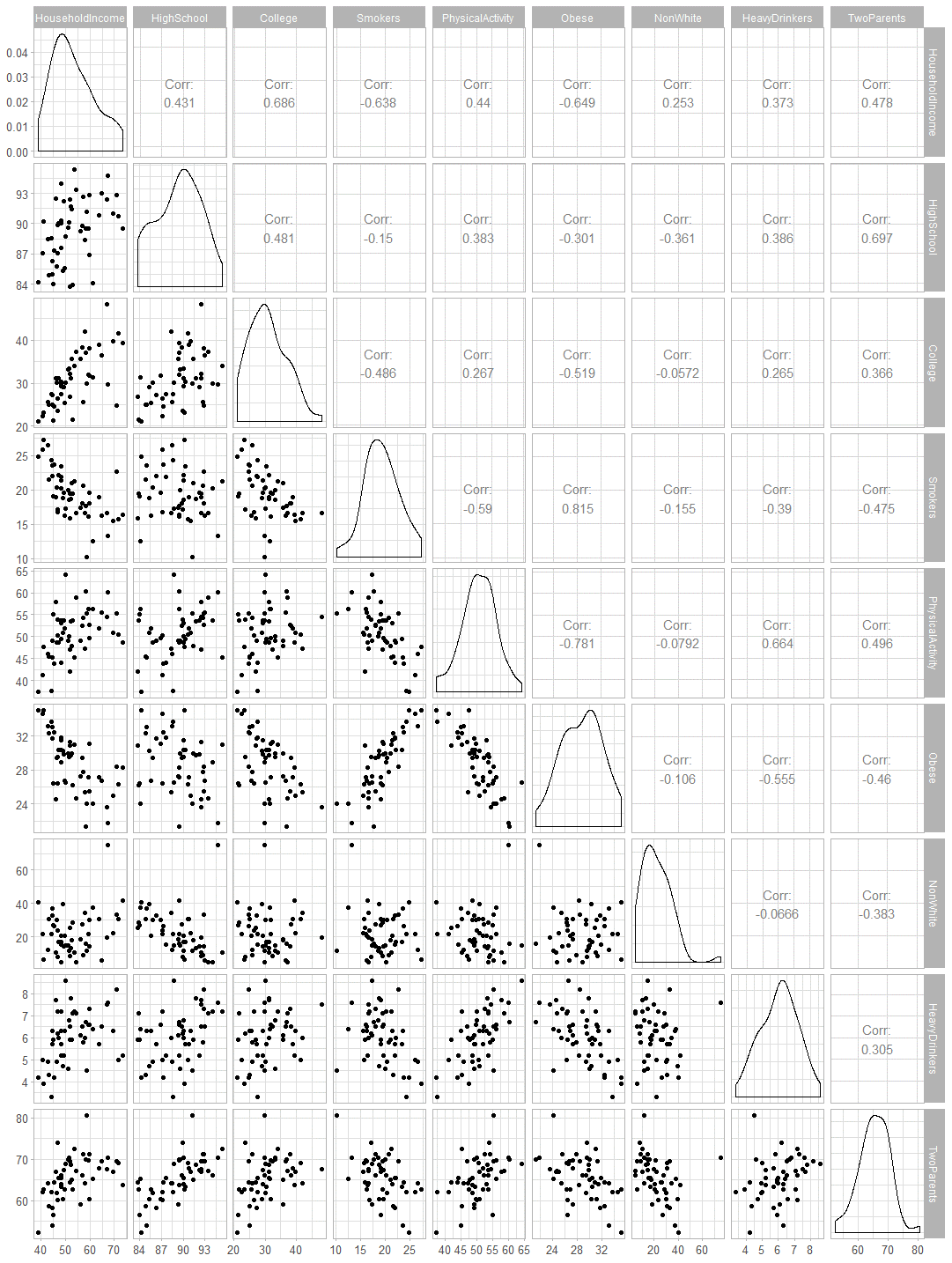
### 2.) For the duration of this assignment, let’s have HOUSEHOLDINCOME be the response variable (Y). Also, please consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain all possible scatterplots relating the non-demographic explanatory variables to the response variable (Y).





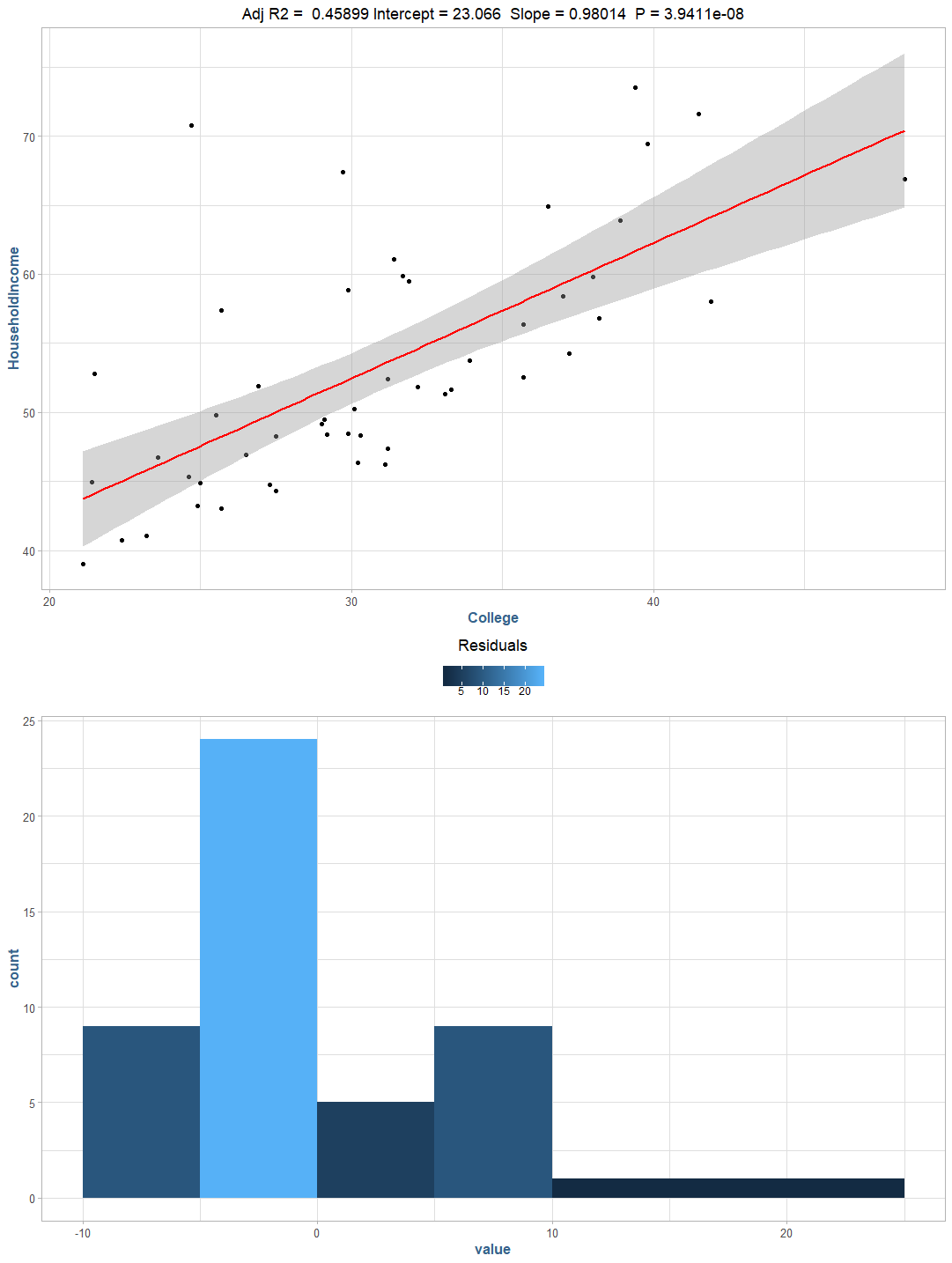
### 3.) Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with Y and report the correlations in a table. Given the scatterplots from step 2) and the correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or why not?





Based upon the correlation to household income, there appears to be three variables which we could fit a linear model to with some success. These variables college, smokers and obese have a semi-colinear relationship to the household income response.

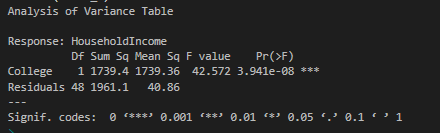
### 4.) Fit a simple linear regression model to predict Y using the COLLEGE explanatory variable. Use the base STAT lm(Y~X) function. Why would you want to start with this explanatory variable? Call this Model 1. Report the results of Model 1 in equation form and interpret each coefficient of the model in the context of this problem. Report the ANOVA table and model fit statistic, R-squared.

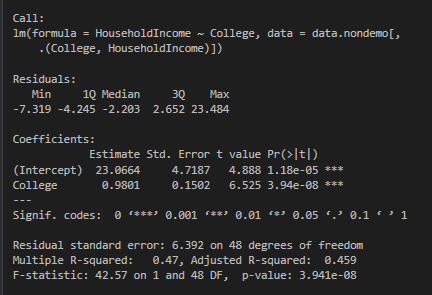


Here, we can see a simple linear model fitted to the college explanatory variable for the household income response variable. We start with the college variable as it has the highest colinearly relationship to the target response variable, household income.

 ŷ = 23.066 + 0.98X1,

where X1 is the percent of resident’s report to have a college education per state.





We can verify the slope and y-intercept by the following long-hand calculations:

slope <- cor(m1$College, m1$HouseholdIncome) \* (sd(m1$HouseholdIncome) / sd(m1$College))

**Output**: 0.9801441

intercept <- mean(m1$HouseholdIncome) - (slope \* mean(m1$College))

**Output:** 23.06644

### 5.) Write R-code to calculate and create a variable of predicted values based on Model 1. Use the predicted values and the original response variable Y to calculate and create a variable of residuals (i.e. residual = Y – Y\_hat = observed minus predicted) for Model 1. Using the original Y variable, the predicted, and/or residual variables, write R-code to:

### Square each of the residuals and then add them up. This is called sum of squared residuals, or sums of squared errors.

m1$Y\_Hat <- predict(model\_1)

m1$residual <- m1$HouseholdIncome - m1$Y\_Hat

sum(m1$residual \*\* 2)

**Output**: 1961.13

### Deviate the mean of the Y’s from the value of Y for each record (i.e. Y – Y\_bar). Square each of the deviations and then add them up. This is called sum of squares total.

y\_bar <- mean(m1$HouseholdIncome)

sum((m1$HouseholdIncome - y\_bar) \*\* 2)

**Output**: 3700.488

### Deviate the mean of the Y’s from the value of predicted (Y\_hat) for each record (i.e. Y\_hat – Y\_bar). Square each of these deviations and then add them up. This is called the sum of squares due to regression.

sum((m1$Y\_Hat - y\_bar) \*\* 2)

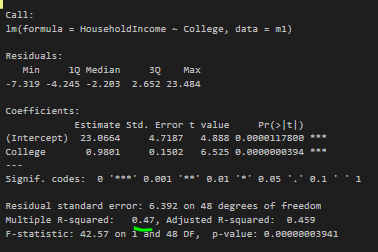
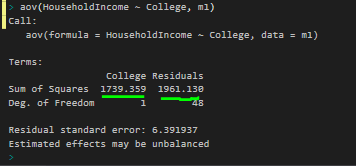
**Output**: 1739.359

### Calculate a statistic that is: (Sum of Squares due to Regression) / (Sum of squares Total)

(ssr / sst)

**Output**: 0.4700349

### Verify and note the accuracy of the ANOVA table and R-squared values from the regression printout from part 4), relative to your computations here.

### 6.)

### 7.)

### 8.)

### 9.)

### 10.)