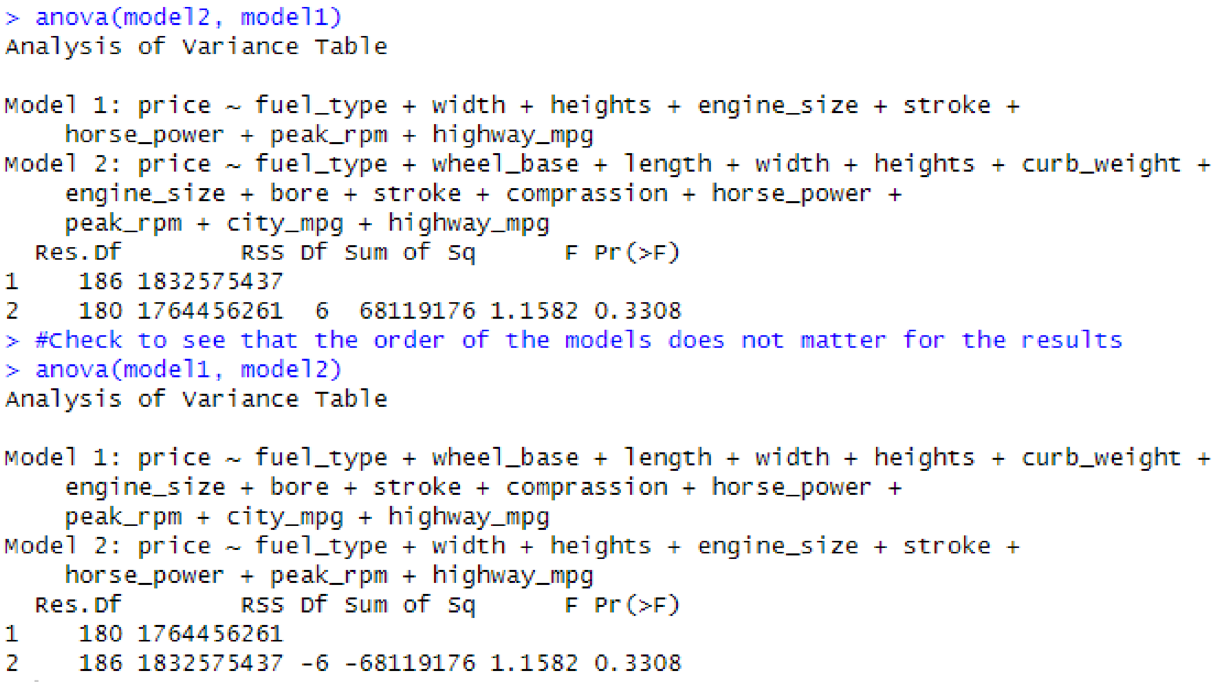
Nghia Lam

1001699317

1. Copy your results for model2 (or model2b): make some observations about your model2 output (Compare them to the full model (model1))
    a. How many variables are significant,
        i. There are 8 Significant figures for model2, all but highway_mpg
        ii. There are 9 Significant figures for model1
    b. What is your adjusted R-squared compared to the full model, in general, do you think your model is better etc.?
        i. Model1 has an R-squared of .849
        ii. Modle2 has an R-squared of .8482
        iii. I believe model2 to be better than model 1 as we have less variables

2. Copy your ANOVA output. What can you tell about the results?

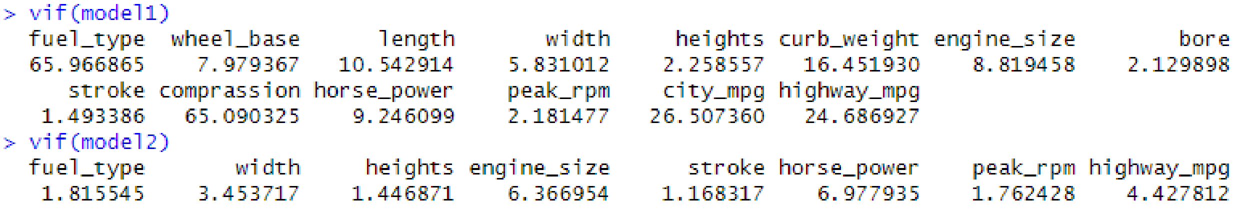


```
> anova(model2, model1)
Analysis of Variance Table

Model 1: price ~ fuel_type + width + heights + engine_size + stroke + 
    horse_power + peak_rpm + highway_mpg
Model 2: price ~ fuel_type + wheel_base + length + width + heights + curb_weight + 
    engine_size + bore + stroke + compression + horse_power + 
    peak_rpm + city_mpg + highway_mpg
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    186 1832575437                             
2    180 1764456261  6  68119176 1.1582 0.3308
> #Check to see that the order of the models does not matter for the results
> anova(model1, model2)
Analysis of Variance Table

Model 1: price ~ fuel_type + wheel_base + length + width + heights + curb_weight + 
    engine_size + bore + stroke + compression + horse_power + 
    peak_rpm + city_mpg + highway_mpg
Model 2: price ~ fuel_type + width + heights + engine_size + stroke + 
    horse_power + peak_rpm + highway_mpg
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    180 1764456261                             
2    186 1832575437 -6 -68119176 1.1582 0.3308
```

3. Compare the full model and reduced model for multicollinearity, what are your observations?

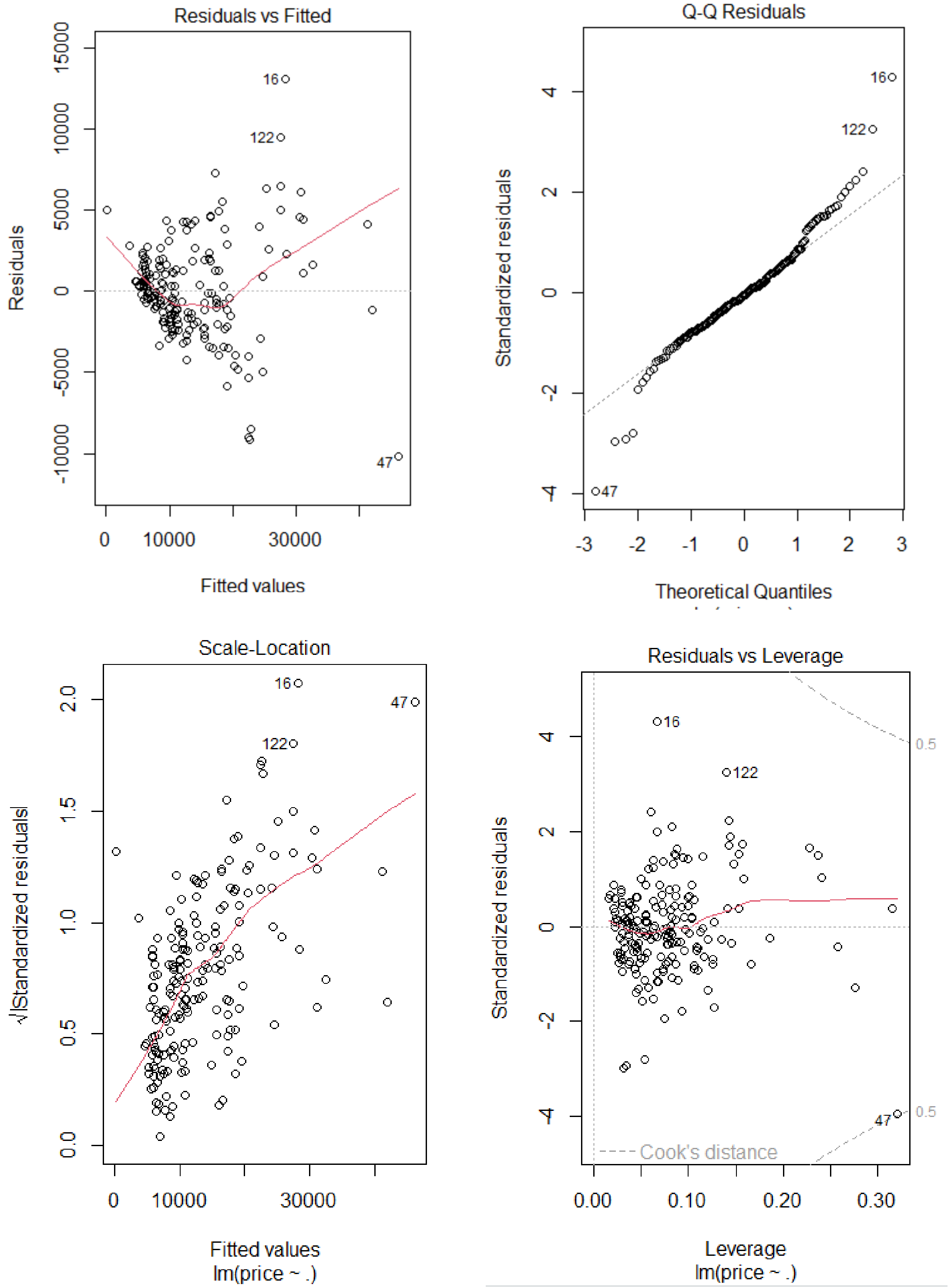


```
> vif(model1)
  fuel_type  wheel_base      length       width     heights curb_weight engine_size        bore 
  65.966865    7.979367   10.542914    5.831012    2.258557   16.451930    8.819458    2.129898 
     stroke compression horse_power    peak_rpm    city_mpg highway_mpg 
   1.493386   65.090325    9.246099    2.181477   26.507360   24.686927 
> vif(model2)
  fuel_type       width     heights engine_size      stroke horse_power    peak_rpm highway_mpg 
   1.815545    3.453717    1.446871    6.366954    1.168317    6.977935    1.762428    4.427812 
```

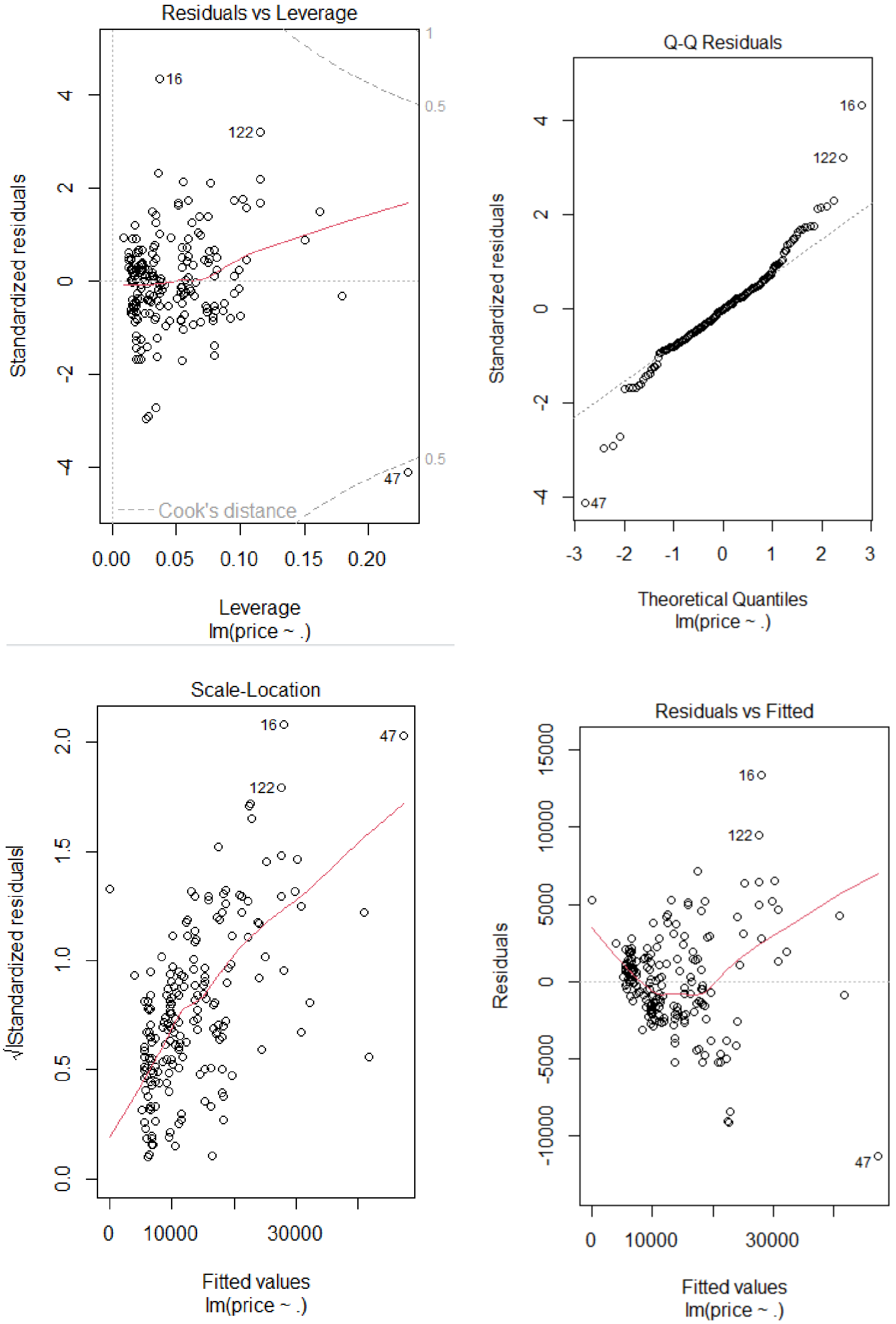
        i. Model 1 has some values with very high multicollinearity like fuel_type, curb_weight, comprassion, city_mpg, highway_mpg.

ii. Model 2 however does not have a multicollinearity problem unlike model1, as all variables are less than 10

4. Copy the plots for the assumptions and compare the results for the reduced and the full model (e.g. did your plots improve?)



i. Above is model1.



ii. Above is model2

iii. Model 1 and model2 look very similar, my plots did not improve.

5. Compare the stepwise selection methods. What did you notice?

    i. Both and Backwards are identical as Model 1 already has all the variables inputted into it so all it would need to do is go backwards, that is why I believe both and backwards are the same.

    ii. Forwards wouldn’t do anything as are the variables are already inputted

    iii. F statistic is overall better for backwards and both, and forwards has a very low F statistic compared to the other 2.

    iv. The p values for all of them were very similar that is why I used it less in my analysis

6. In general, comparing the process for regression analysis in Python and R, which one did you like better and why?

    i. I liked R because it was easier to implement everything compared to python.