Solution to the Midterm Assignment (401)
Part A

Group Number: 5
Group Member 1 (Name): Kirtivardhan Singh
Group Member 1 (Exam number): 20227707131

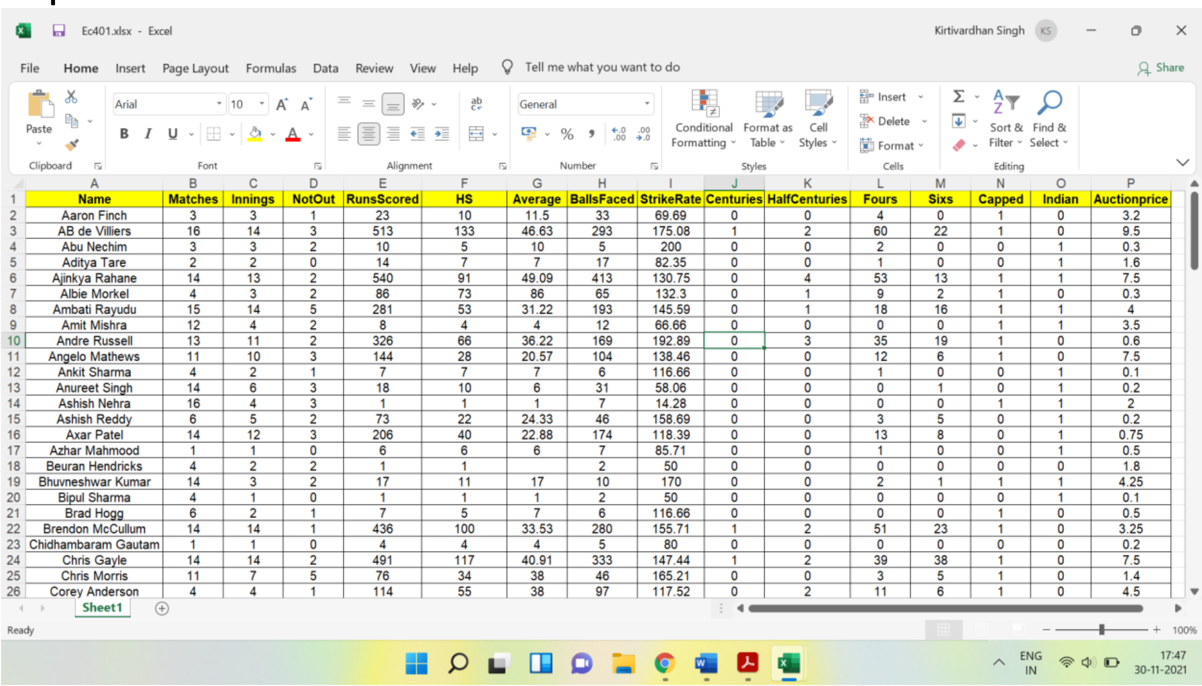
Group Member 2 (Name): Bhanvi Kakkar
Group Member 2 (Exam number): 20227707043

Question 1: ***Year-2015***

- ***Provide your final data. This can be in excel/stata/r format. I should be able to read this data:***

- ***Provide your program. I should be able to run the program one-click using your data. Provide sufficient comments within the program file so that I can understand what you are doing there:***

Snapshot of Data:



Name	Matches	Innings	NotOut	RunsScored	HS	Average	BallsFaced	StrikeRate	Centuries	HalfCenturies	Fours	Sixs	Capped	Indian	Auctionprice
Aaron Finch	3	3	1	23	10	11.5	33	69.69	0	0	4	0	1	0	3.2
AB de Villiers	16	14	3	513	133	46.63	293	175.08	1	2	60	22	1	0	9.5
Abu Nechim	3	3	2	10	5	10	5	200	0	0	2	0	0	1	0.3
Aditya Tare	2	2	0	14	7	7	17	82.35	0	0	1	0	0	1	1.6
Ajinkya Rahane	14	13	2	540	91	49.09	413	130.75	0	4	53	13	1	1	7.5
Albie Morkel	4	3	2	86	73	86	65	132.3	0	1	9	2	1	0	0.3
Ambati Rayudu	15	14	5	281	53	31.22	193	145.59	0	1	18	16	1	1	4
Amit Mishra	12	4	2	8	4	4	12	66.66	0	0	0	0	1	1	3.5
Andre Russell	13	11	2	326	66	36.22	169	192.89	0	3	35	19	1	0	0.6
Angelo Mathews	11	10	3	144	28	20.57	104	138.46	0	0	12	6	1	0	7.5
Ankit Sharma	4	2	1	7	7	7	6	116.66	0	0	1	0	0	1	0.1
Anureet Singh	14	6	3	18	10	6	31	58.06	0	0	0	1	0	1	0.2
Ashish Nehra	16	4	3	1	1	1	7	14.28	0	0	0	0	1	1	2
Ashish Reddy	6	5	2	73	22	24.33	46	158.69	0	0	3	5	0	1	0.2
Axar Patel	14	12	3	206	40	22.88	174	118.39	0	0	13	8	0	1	0.75
Azhar Mahmood	1	1	0	6	6	6	7	85.71	0	0	1	0	0	1	0.5
Beuran Hendricks	4	2	2	1	1		2	50	0	0	0	0	0	0	1.8
Bhuvneshwar Kumar	14	3	2	17	11	17	10	170	0	0	2	1	1	1	4.25
Bipul Sharma	4	1	0	1	1	1	2	50	0	0	0	0	0	1	0.1
Brad Hogg	6	2	1	7	5	7	6	116.66	0	0	0	0	1	0	0.5
Brendon McCullum	14	14	1	436	100	33.53	280	155.71	1	2	51	23	1	0	3.25
Chidhambaram Gautam	1	1	0	4	4	4	5	80	0	0	0	0	0	0	0.2
Chris Gayle	14	14	2	491	117	40.91	333	147.44	1	2	39	38	1	0	7.5
Chris Morris	11	7	5	76	34	38	46	165.21	0	0	3	5	1	0	1.4
Corey Anderson	4	4	1	114	55	38	97	117.52	0	2	11	6	1	0	4.5

Steps Followed:

1. The above data is inputted in the Python code as a Data Frame
2. There are few missing values in average. They are imputed by 'Mean of Averages' of other players. Mean is used to replace missing values in the above data
3. Firstly, we ran regression of all the variables shown above using the regression equation:

$$Auctionprice_i = \alpha_i + \beta_1 X_{i1} + \cdots . \beta_k X_{ik} + \mu_i$$

$$where\ i = 1,2, \ldots 129\ and\ k = 1,2 \ldots 14$$

4. The two variables we introduced in our regression that is different from the data source are:

Dependent Variable: Auction Price

Variable Name	**Variable Description**
Capped	Introduced as a dummy variable i.e. 0 for uncapped (having no international cricket experience and 1 for capped player) as of 2015 IPL.
Indian	If the player is not a citizen of India- 0, Otherwise- 1

5. Before running the regression, our ex-ante expectation of the sign of the beta coefficients and the reasons for such expectations were:

Variable Name	**A priori Expected Sign**	**Reason**
Matches	+	Higher Number of matches-In form batsman
Innings	+	More innings, more batting opportunities, preferably top/middle order batsman
NotOut	+	Better scoring opportunities
RunsScored	+	Better asset to team
HS	+	Leading runs scorer, better average and total runs
Average	+	Higher Average, better batsman
BallsFaced	Ambiguous	More number of balls means more scoring opportunities and higher runs, but can have lower strike rate
StrikeRate	+	Finisher, hard hitter of ball hence favourable to any team
Centuries	+	More Centuries, more reliable batsman
HalfCenturies	+	More half centuries, more reliable batsman
Fours	+	More fours, more runs
Sixs	+	More sixes, more runs

Capped	+	Capped batsman-more valuable, more experience
Indian	Ambiguous	You can have a total of as many Indians as you want in a game contrary to upper cap of 4 International players, but then international players since they are limited, teams strive to have best international players.

```
==============================================================================
Dep. Variable:           Auctionprice   R-squared:                       0.488
Model:                            OLS   Adj. R-squared:                  0.425
Method:                 Least Squares   F-statistic:                     7.761
Date:                Tue, 30 Nov 2021   Prob (F-statistic):           2.75e-11
Time:                        12:28:38   Log-Likelihood:                -303.67
No. Observations:                 129   AIC:                             637.3
Df Residuals:                     114   BIC:                             680.2
Df Model:                          14
Covariance Type:            nonrobust
=================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            0.4520      0.981      0.461      0.646      -1.491       2.395
Matches          0.0808      0.089      0.913      0.363      -0.095       0.256
Innings          0.1193      0.218      0.546      0.586      -0.313       0.552
NotOut          -0.0723      0.255     -0.283      0.778      -0.578       0.433
RunsScored       0.0058      0.022      0.264      0.792      -0.038       0.049
HS              -0.0100      0.025     -0.410      0.683      -0.059       0.039
Average         -0.0016      0.035     -0.046      0.963      -0.071       0.068
BallsFaced       0.0159      0.019      0.816      0.416      -0.023       0.055
StrikeRate      -0.0076      0.007     -1.067      0.288      -0.022       0.007
Centuries        3.8540      2.554      1.509      0.134      -1.206       8.914
HalfCenturies    0.0067      0.524      0.013      0.990      -1.031       1.044
Fours           -0.0958      0.090     -1.065      0.289      -0.274       0.082
Sixs            -0.0771      0.128     -0.605      0.546      -0.330       0.175
Capped           2.8855      0.612      4.713      0.000       1.673       4.098
Indian           0.1542      0.593      0.260      0.795      -1.020       1.328
==============================================================================
```

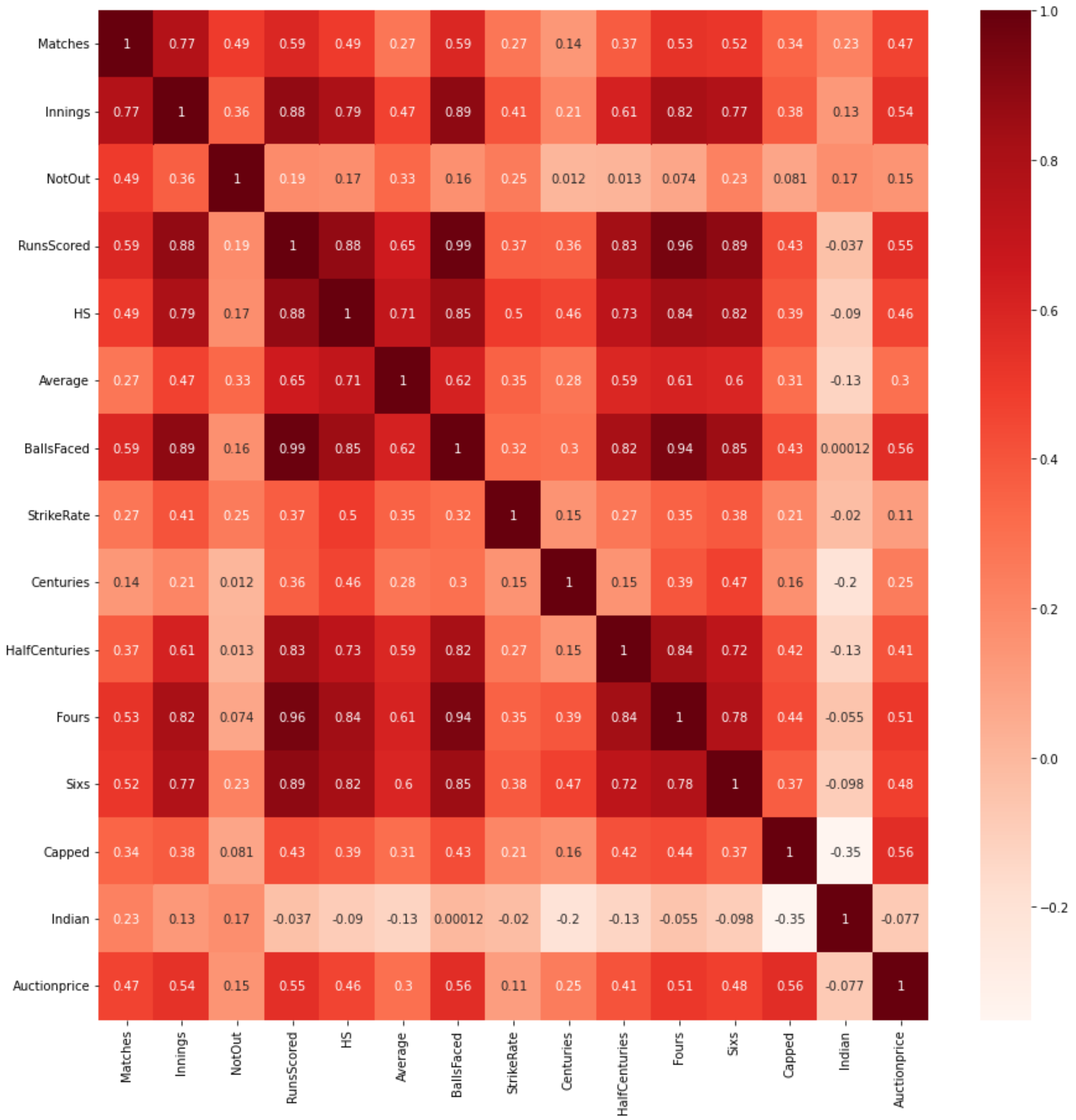
6. It is evident from the above snapshot that the results are not very encouraging. **R square is decent but t ratios are insignificant**. This suggests presence of **Multicollinearity** in the data.
   The only variable that was found to be significant at this stage was dummy variable introduced by us- *Capped* meaning that whether the player has international match experience does affect its auction price.

7. At next step, we checked for Multicollinearity between the independent variables and with the dependent variable.

We also employed **Feature Selection Method** to improve performance of the model and to avoid curse of dimensionality.

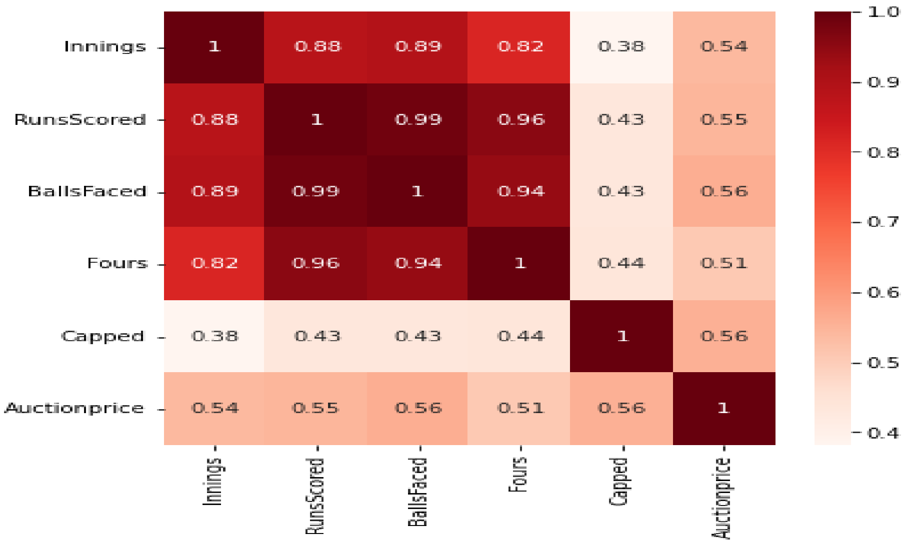
The feature selection method employed was ***Filter Method*** based on the fact that if an independent variable has **correlation >0.5** with the dependent variable, it will be added to the regression equation provided it does not have very high correlation with some other independent variable.

Here's the result of Feature Selection Method (Filter Method)



8. Based on the above Heat Map (based on Pearson Correlation Coefficient), only 5 variables are selected- *Innings, RunScored, Ballsfaced, Fours and Capped.*

9. Although the filter method suggested taking all 5 variables, but one of the important assumptions of Ordinary Least Squares (OLS) is not to have perfect multicollinearity. Hence, we checked correlation between these five variables.



10. There is near perfect Multicollinearity between RunsScored and BallsFaced and hence, we dropped BallsFaced from our Regression Equation.

We have also estimated Regression Equation with RunsScored as Independent Variable but due to its high correlation with Fours and Innings, fewer ratios were coming out significant, but we still kept it to avoid ***Omitted Variable Bias.***

Although we have reported Regression results with RunsScored as well, we didn't consider it in final variable as dropping it led to a significant value of Innings and also because Innings are less correlated with fours than runs scored. So we could keep both Innings and Fours in our Regression analysis by dropping RunsScored.

11. The final set of Independent Variables and their expected signs are:

Variable Name	A priori Expected Sign
Innings	+
Fours	+
Capped	+

Null Hypothesis

H0: Innings doesn't have a significant effect on Auction Price
H0: Fours doesn't have a significant effect on Auction Price
H0: Player being capped doesn't have a significant effect on Auction Price

HA:. Innings have a positive significant effect on Auction Price
H0: Fours have a significant effect on Auction Price
H0: Player being capped have a significant effect on Auction Price

12. Since the number of observations (129) are large enough (>30) by Central Limit theorem, Z-test can be applied and the critical value at 1%, 5% and 10% are 2.33, 1.64 and 1.28 for a **one tailed test.**

    In the final regression output given below, the signs of all the three variables are as expected. *Innings* and *Capped* are significant even at **1% level of significance** implying that they will be significant even at 5% or 10% significance level.

    However, we could not find much evidence to suggest that *Fours* has significant effect on auction price.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Auctionprice   R-squared:                       0.436
Model:                            OLS   Adj. R-squared:                  0.423
Method:                 Least Squares   F-statistic:                     32.23
Date:                Tue, 30 Nov 2021   Prob (F-statistic):           1.68e-15
Time:                        14:52:32   Log-Likelihood:                -309.89
No. Observations:                 129   AIC:                             627.8
Df Residuals:                     125   BIC:                             639.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2075      0.481     -0.431      0.667      -1.160       0.745
Innings        0.2482      0.085      2.925      0.004       0.080       0.416
Fours          0.0128      0.027      0.468      0.641      -0.041       0.067
Capped         2.8868      0.535      5.394      0.000       1.828       3.946
==============================================================================
Omnibus:                       20.738   Durbin-Watson:                   2.174
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.087
```

<u>Inference</u>

Null hypothesis is rejected for *Innings* and *Capped* Independent **Variables in favour of Alternate Hypothesis.**

This imply that a greater number of Innings a batsman gets to play means he's an important asset to the team and his batting performance is good therefore he's either a top order batsman or a finisher and his coming to bat significantly improves team performance and justify the high auction price for him. That's why significance at 1% level.

Capped is also significant even at 1% level, suggesting that if a player has made his international debut, he may fetch more amount of auction price than a person who has just made appearances in domestic series. This seems natural as only the best of domestic players gets to represent their nations at international level.

Q1 Part B

13. For Part B of the question, we wrote code to fetch the first name of players. We even made manual corrections such as for AB Deviliers, his first name is Abraham.

14. Once the first name is fetched, we took the codes from GitHub that uses ASCII letters to convert Names to numbers. The details of the code is shared and proper comments have been introduced in the code.
15. Num is the variable that stores the numerology values. Now we ran two types of regression- Simple and Multiple Regression.

16. Although, we consider num to be an irrelevant variable, still from what we can infer from the question, based on the belief, a higher value in numerology signifies a higher auction price and hence we expect the sign of beta coefficient to be **positive.**

    H0: num has no effect on Auction price
    HA: num has a positive effect on Auction price

    Now the Regression Results:

    Simple Linear Regression:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Auctionprice   R-squared:                       0.025
Model:                            OLS   Adj. R-squared:                  0.017
Method:                 Least Squares   F-statistic:                     3.244
Date:                Tue, 30 Nov 2021   Prob (F-statistic):             0.0741
Time:                        14:52:32   Log-Likelihood:                -345.22
No. Observations:                 129   AIC:                             694.4
Df Residuals:                     127   BIC:                             700.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.2170      0.663      3.345      0.001       0.906       3.529
x1             0.2153      0.120      1.801      0.074      -0.021       0.452
==============================================================================
Omnibus:                       37.871   Durbin-Watson:                   2.092
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               59.824
Skew:                           1.476   Prob(JB):                     1.02e-13
Kurtosis:                       4.555   Cond. No.                         12.1
==============================================================================
```

Since the number of observations (129) are large enough (>30) by Central Limit theorem, Z-test can be applied and the critical value at 1%, 5% and 10% are 2.33, 1.64 and 1.28 for a one tailed test.

But since, the t-value is 1.80, it is significant only at 10% hence weakly significant. This significant we expect to go as relevant variables are added.

Multiple Linear Regression Result

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Auctionprice   R-squared:                       0.441
Model:                            OLS   Adj. R-squared:                  0.423
Method:                 Least Squares   F-statistic:                     24.45
Date:                Tue, 30 Nov 2021   Prob (F-statistic):           6.23e-15
Time:                        14:52:32   Log-Likelihood:                -309.34
No. Observations:                 129   AIC:                             628.7
Df Residuals:                     124   BIC:                             643.0
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.6080      0.618     -0.983      0.327      -1.832       0.616
Innings        0.2360      0.086      2.756      0.007       0.067       0.406
Fours          0.0138      0.027      0.504      0.615      -0.040       0.068
Capped         2.8925      0.535      5.405      0.000       1.833       3.952
num            0.0960      0.093      1.030      0.305      -0.088       0.281
==============================================================================
Omnibus:                       21.294   Durbin-Watson:                   2.136
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.375
Skew:                           0.927   Prob(JB):                     1.14e-06
Kurtosis:                       4.287   Cond. No.                         60.7
==============================================================================
```

Since the number of observations (129) are large enough (>30) by Central Limit theorem, Z-test can be applied and the critical value at 1%, 5% and 10% are 2.33, 1.64 and 1.28 for a one tailed test.

But since, the t-value is 1.030, num is not significant at even 10% level. Innings and Capped are still significant even at 1% level of significance.

Hence auction price is significantly explained by Innings and Capped and inclusion of irrelevant variable doesn't affect output values much.

Proper Explanation using Matrices:

Impact of inclusion of the variable on earlier coefficient

Let our initial model be

$$y = X \cdot b + e$$

Now on adding a variable, the new model becomes

$$y = X d + Z c + u$$

The normal equations are

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix} \begin{bmatrix} d \\ c \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

$$X'X\,d + X'Z\,c = X'y$$
$$d = (X'X)^{-1}(X'y - X'Zc)$$
$$= (X'X)^{-1}X'(y - Zc)$$

This is not equal to $b = (X'X)^{-1}X'y$ from the initial regression

We know $y \geq y - Zc$ so coefficients in b will be larger than those in d. Thus inclusion of numerology variable reduces value of existing beta coefficients.

Using Frisch - Waugh - Lovell Theorem,

$$c = \frac{z^{*\prime} y^*}{z^{*\prime} z^*}$$

$$u = y - Xb + X(X'X)^{-1}X'Zc - Zc$$
$$= e - M_X Zc$$
$$= e - z^* c$$

$$u'u = e'e + c^2 z^{*\prime}z^* - 2c(z^{*\prime}e)$$
$$= e'e - c^2(z^{*\prime}z^*)$$
$$\leq e'e$$

$\therefore$ sum of squares will never decrease due to inclusion of a variable so R^2 will also never decrease

Also by Frisch - Waugh - Lovell Th^m

$$d = (X' M_Z X)^{-1} (X' M_Z y)$$

$$= (X' M_Z X)^{-1} (X' M_Z (Xb + e))$$

$$= (X' M_Z X)^{-1} (X' M_Z X) b + (X' M_Z X)^{-1} X' M_Z e$$

$$= b + (X' M_Z X)^{-1} X' M_Z e$$

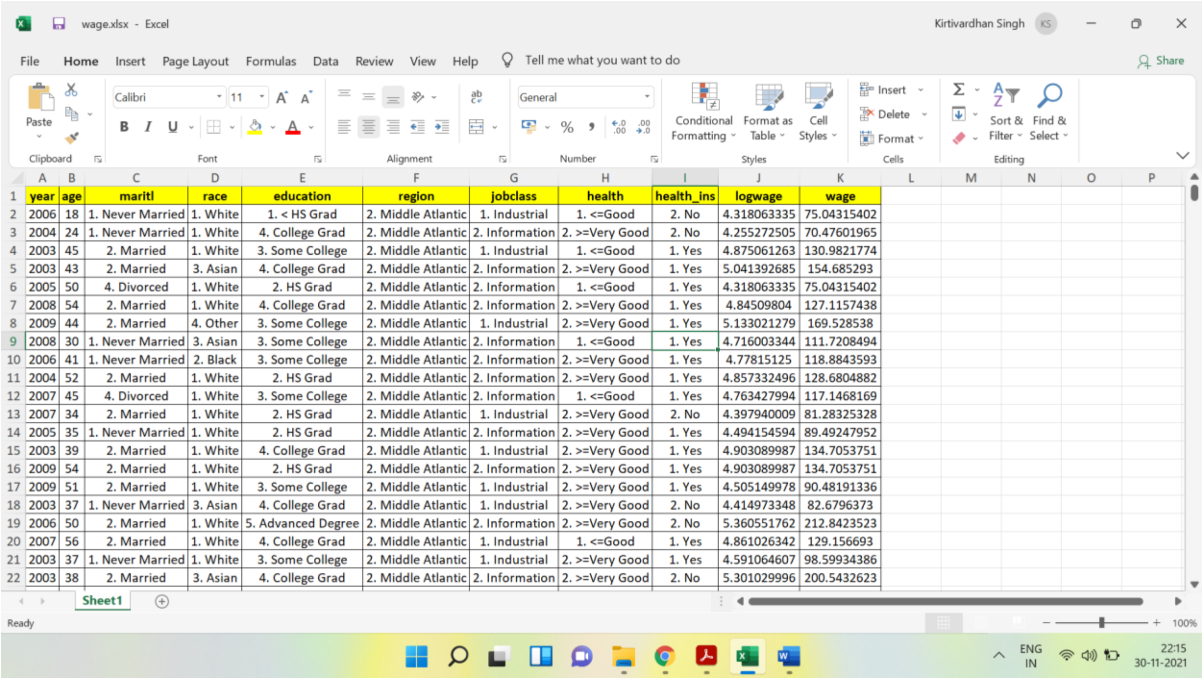
$$E(d) = E(b)$$

$\therefore$ The least squares estimator is unbiased even after inclusion of irrelevant variable, but the cost of overspecifying the model is larger variances of the estimators.

Question 2:

X variable- Jobclass

1. Snapshot of Data



	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
2	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063335	75.04315402
3	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255272505	70.47601965
4	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061263	130.9821774
5	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041392685	154.685293
6	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063335	75.04315402
7	2008	54	2. Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.84509804	127.1157438
8	2009	44	2. Married	4. Other	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.133021279	169.528538
9	2008	30	1. Never Married	3. Asian	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.716003344	111.7208494
10	2006	41	1. Never Married	2. Black	3. Some College	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.77815125	118.8843593
11	2004	52	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.857332496	128.6804882
12	2007	45	4. Divorced	1. White	3. Some College	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.763427994	117.1468169
13	2007	34	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.397940009	81.28325328
14	2005	35	1. Never Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.494154594	89.49247952
15	2003	39	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.903089987	134.7053751
16	2009	54	2. Married	1. White	2. HS Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	4.903089987	134.7053751
17	2009	51	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.505149978	90.48191336
18	2003	37	1. Never Married	3. Asian	4. College Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.414973348	82.6796373
19	2006	50	2. Married	1. White	5. Advanced Degree	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	5.360551762	212.8423523
20	2007	56	2. Married	1. White	4. College Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.861026342	129.156693
21	2003	37	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.591064607	98.59934386
22	2003	38	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	5.301029996	200.5432623

2. Firstly, data cleaning is done as the data had a lot of variables that were not relevant to us. These variables are dropped and in subsequent steps data is cleaned.

3. Since our data has 1 Quantitative/Numerical Variable (Age) and 2 Qualitative/Categorical variables (education and jobclass).

4. To deal with categorical variables, we need to convert them to dummy variables, since our regression will have a intercept, we need to introduce n-1 categories of dummies for a dummy variable having n categories to avoid Dummy Variable Trap.

5. Since education has 5 categories, 4 dummies are introduced and since jobclass has 2 categories, one dummy is introduced for it. One Hot Encoding and Label Encoding are two popular methods to dead with categorical variables but since our variable of interest- jobclass have just 2 categories- The difference between the two is irrelevant to us and we used the following nomenclature:

Base Category for Education: < HS Grad
Base Category for jobclass: Industrial

The dummy variables are introduced for the rest categories of Education and jobclass.

6. The major difference between a normal regression involving just Quantitative variables vs one Involving Categorical variables are that the intercept represents the values for base category and the Beta coefficients are **differential slope coefficients**. Both Qualitative and Quantitative variables are important to data and there's no one better than the other. Both tries to explain the Dependent variable.

$$logwage_i = \beta_o + \beta_1 age_i + \beta_2 education_i + \beta_3 X + \mu_i$$

7. Our Ex-ante Expectation was:

Variable Name	A priori Expected Sign	Reason
intercept	+	Wage can't be negative
Age	+	Wages on average should increase with age keeping other factors constant
Education dummies	+	Higher the education level, higher the wage on average. Since the basic level of education is taken as base class, the differential slopes should be positive
Jobclass dummies	+	Information workers in general have more pay than industrial workers. Industrial worker is taken as dummy.

<u>Null Hypothesis</u>

H0: No effect of Age on Wages
H0: No effect of Experience on Wages
H0: No effect of Jobclass on Wages

HA: Positive effect of ageing on Wages
HA: Positive effect of more education on Wages
HA: Positive effect of Information Worker on Wages

Since the number of observations are large enough (>30) by Central Limit theorem, Z-test can be applied and the critical value at 1%, 5% and 10% are 2.33, 1.64 and 1.28 for a one tailed test.

Regression Output

```
==============================================================================
Dep. Variable:                logwage   R-squared:                       0.262
Model:                            OLS   Adj. R-squared:                  0.260
Method:                 Least Squares   F-statistic:                     176.9
Date:                Tue, 30 Nov 2021   Prob (F-statistic):          4.10e-193
Time:                        16:28:40   Log-Likelihood:                -666.53
No. Observations:                3000   AIC:                             1347.
Df Residuals:                    2993   BIC:                             1389.
Df Model:                           6
Covariance Type:            nonrobust
=============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                      4.1583      0.027    151.829      0.000       4.105       4.212
age                        0.0055      0.000     11.325      0.000       0.005       0.006
educ_ Advanced Degree      0.5252      0.024     21.671      0.000       0.478       0.573
educ_ College Grad         0.3568      0.022     16.152      0.000       0.313       0.400
educ_ HS Grad              0.1186      0.021      5.680      0.000       0.078       0.160
educ_ Some College         0.2363      0.022     10.707      0.000       0.193       0.280
jobclass_ Information      0.0374      0.012      3.216      0.001       0.015       0.060
==============================================================================
Omnibus:                      319.795   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1158.631
Skew:                          -0.499   Prob(JB):                    2.55e-252
Kurtosis:                       5.876   Cond. No.                         349.
==============================================================================
```

8. Since the t coefficients are higher than 2.33, all the independent variables are significant at 1% level and Null hypothesis is rejected.

9. An industrial jobclass worker who <HS graduate has a coefficient of 4.1583, whereas a HS graduate Industrial worker earns 11.86% higher than an industrial jobclass worker who <HS graduate. Industrial worker who has advance degree earns 52.52% higher than an industrial jobclass worker who <HS graduate. Also, an information jobclass worker earn 3.74% more than a industrial jobclass worker of same education level.

10. We can conclude that Higher Age on average leads to higher wages. Higher Education level in general leads to higher wages and Information jobclass on average earns higher than Industrial jobclass.

Part B

Two Regressions
1. $age_i = \beta_1 + \beta_2 Xi$ and residuals are stored as age*
2. $education_i = \beta_3 + \beta_4 Xi$ and residuals are stored as education*

There was just 1 regression for age* but 4 regressions ran for education* and then finally using the equation

$$logwage_i = \alpha_o + \alpha_1 age_i * + \alpha_2 education_i * + \Omega_i$$

Dep. Variable:	logwage	R-squared:	0.220
Model:	OLS	Adj. R-squared:	0.218
Method:	Least Squares	F-statistic:	168.5
Date:	Tue, 30 Nov 2021	Prob (F-statistic):	2.81e-158
Time:	16:28:40	Log-Likelihood:	-749.90
No. Observations:	3000	AIC:	1512.
Df Residuals:	2994	BIC:	1548.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6539	0.006	819.636	0.000	4.643	4.665
agestar	0.0055	0.000	11.017	0.000	0.004	0.006
educ1star	0.5252	0.025	21.080	0.000	0.476	0.574
educ2star	0.3568	0.023	15.712	0.000	0.312	0.401
educ3star	0.1186	0.021	5.525	0.000	0.077	0.161
educ4star	0.2363	0.023	10.415	0.000	0.192	0.281

Omnibus:	279.006	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	973.608
Skew:	-0.435	Prob(JB):	3.83e-212
Kurtosis:	5.652	Cond. No.	81.9

Which has same Differential slope coefficients as the Initial Regression Equation.

Possible Explanation using Matrices

We start with our initial model

$$\log wage_i = \beta_0 + \beta_1 age_i + \beta_2 education_i + \beta_3 X + \epsilon_i$$

this can be written as

$$y = X_1 B_1 + X_2 B_2 + \epsilon$$

where $y = (n \times 1)$ matrix on observations of log wage

$X_1 = (n \times 3)$ matrix with a column of 1s and observations regarding age, & education

$X_2 = (n \times 1)$ matrix of observations of job - class

$x_1 = [1 \quad age \quad education] \qquad x_2 = [$job - cla

$x = [x_1 \quad x_2]$

Normal eq^n: $(x'x)\,b = x'y$

$$\begin{bmatrix} x_1'x_1 & x_1'x_2 \\ x_2'x_1 & x_2'x_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} x_1'y \\ x_2'y \end{bmatrix}$$

$$x_1'x_1b_1 + x_1'x_2b_2 = x_1'y$$

$$x_1'x_1b_1 = x_1'y - x_1'x_2b_2$$

$$b_1 = (x_1'x_1)^{-1}[x_1'y - x_1'x_2b_2]$$

$$= (x_1'x_1)^{-1}(x_1'y) - (x_1'x_1)^{-1}(x_1'x_2b_2)$$

$$= (x_1'x_1)^{-1}x_1'(y - x_2b_2)$$

2. $X_2' X_1 b_1 + X_2' X_2 b_2 = X_2' y$

$X_2' X_2 b_2 = X_2' y - X_2' X_1 b_1$

$= X_2' (y - X_1 b_1)$

$b_2 = (X_2' X_2)^{-1} X_2' (y - X_1 b_1)$

Using b_1 & b_2 in 2.

$X_2' X_1 (X_1' X_1)^{-1} X_1' (y - X_2 b_2) + X_2' X_2 b_2 = X_2' y$

$\rightarrow X_2' X_1 (X_1' X_1)^{-1} X_1' y - X_2' X_1 (X_1' X_1)^{-1} X_1' X_2 b_2 + X_2' X_2 b_2 = X_2' y$

$\rightarrow$ Collecting & re-arranging terms

$X_2' [I - X_1 (X_1' X_1)^{-1} X_1'] X_2 b_2 = X_2' [I - X_1 (X_1' X_1)^{-1} X_1'] y$

$\rightarrow X_2' M_1 X_2 b_2 = X_2' M_1 y$

$\rightarrow b_2 = (X_2' M_1 X_2)^{-1} (X_2' M_1 y)$

where $M_1 = I - X_1 (X_1' X_1)^{-1} X_1'$ is residual maker for a regression on columns of X_1

$\rightarrow$ Similarly $b_1 = (X_1' M_2 X_1)^{-1} (X_1' M_2 y)$

where $M_2 = I - X_2 (X_2' X_2)^{-1} X_2'$

is residual maker for regression on X_2

since M is symmetric & idempotent, so is M_1 & M_2

$\Rightarrow b_1 = (X_1' M_2' M_2 X_1)^{-1} (X_1' M_2 y)$

$= (X_1^{*'} X_1^*)^{-1} (X_1^{*'} y)$

This is same ~~a~~ as coeffecient vector in regression of log wage on age* and education*

Here $X_1^* = M_2 X_1$ = Residuals obtained when X_1 regressed on X_2

$\therefore$ Coefficients in both models are same for age and education.