Exploratory Data Analysis on "blood pressure data"

Sisay Menji Bekena

In [106]:
```
# import necessary libraries
import pandas as pd, numpy as np, matplotlib.pyplot as plt
import scipy as sp, statsmodels.api as sm, researchpy as rp
from sklearn.preprocessing import scale
from scipy import stats
from statsmodels.formula.api import ols

%matplotlib inline
```

Data Description

The data is taken from Stata statistical software fictional blood pressure data. It can be acquired within stata by typing the command **sysuse bplong**. The data contains paired sample observation of a blood pressure experiment and has five columns as follows:

- patient: the id of the patient
- sex: binary variable: Male or Female
- agegrp: the age group of the patient
- when: a binary variable showing whether the blood pressure measurement is before or after the experiment
- bp: the blood pressure measurement

Objectives of the Exploratory Data Analysis (EDA)

In this project I want to explore the data to understand the following points

1. Understand whether blood pressure varies across gender and age group
2. Test whether significant differences exist in blood pressure between male and female
3. Conduct anova test t understand the combined effect of age group and sex on blood pressure
4. Test whether the experiment affected the after blood pressure measurement

I will use only the before data sets for the questions in in 1-3 and will use the data including the after measurement for question 4. Since the data is not in the format required I will do data transformation next.

Data preparation

In [107]:
```
df = pd.read_excel("bplong.xlsx", names=["patient","sex","agegrp","when","bp"]
)
```

In [108]:
```
#long data set shape
df.shape, df.columns
```

Out[108]:
```
((239, 5), Index(['patient', 'sex', 'agegrp', 'when', 'bp'], dtype='object'))
```

In [109]:
```
df.head()
```

Out[109]:

	patient	sex	agegrp	when	bp
0	1	Male	30-45	After	153
1	2	Male	30-45	Before	163
2	2	Male	30-45	After	170
3	3	Male	30-45	Before	153
4	3	Male	30-45	After	168

As seen above the data is in long format where the "when" variable should have been two columns separately for the before and after measurements. I will change the data into wide format next conducting the following steps:

- change the data into wide using "patient id" as the index value and "when" as the column identifier for blood pressure (bp)
- remove the extra columns created for sex and agegrp for the two values of bp
- drop observations where the agegrp, sex or bp measurement are missing

In [110]:
```
# idex the data by the categorical variables: will not change before and after
df_wide = df.pivot(index="patient", columns='when', values=["bp","sex", "agegr
p"])
wide_cols = list(df_wide.columns)
rename_cols = [x[0].lower()+"_"+ x[1].lower() for x in wide_cols]
# rename the columns to bp_after syntax and drop the before values for sex and
agegrp
df_wide.columns = rename_cols
df_wide.columns
```

Out[110]:
```
Index(['bp_after', 'bp_before', 'sex_after', 'sex_before', 'agegrp_after',
       'agegrp_before'],
      dtype='object')
```

In [111]:
```
# drop extra columns of sex and agegrp
df_wide.drop("sex_after", axis=1, inplace=True)
df_wide.drop("agegrp_after", axis=1, inplace=True)
# rename back
df_wide.reset_index(inplace=True)
df_wide.columns = ["patient","bp_after","bp_before","sex","agegrp"]
print(df_wide.shape)
df_wide.head()
```

```
(120, 5)
```

Out[111]:

	patient	bp_after	bp_before	sex	agegrp
0	1	153	NaN	NaN	NaN
1	2	170	163	Male	30-45
2	3	168	153	Male	30-45
3	4	142	153	Male	30-45
4	5	141	146	Male	30-45

In [112]:
```
# drop null values
df_wide.dropna(axis=0, how='any',inplace=True)
df_wide["bp_after"] = df_wide["bp_after"].astype("int")
df_wide["bp_before"] = df_wide["bp_before"].astype("int")
df_wide.shape
```

Out[112]:
```
(119, 5)
```

In [113]:
```
# check data types
df_wide.dtypes
```

Out[113]:
```
patient       int64
bp_after      int32
bp_before     int32
sex          object
agegrp       object
dtype: object
```

As can be seen above dropping one observation for patient 1 with null values reduces the number of observations by 1

Descriptive analysis

After exploring the distribution of the individual variables below, I will explore the interaction of age group and sex with gender before going for formal hypothesis testing. The tables below show that:

- **blood pressure** is a numeric variable
- **age group** is a categorical variable with three options, 30-45, 46-59 and 60+ with nearly equal distribution between the age groups
- **sex** is a categorical variable with 120 female and 119 male observations

In [114]:
```
df_wide["bp_before"].describe()
```

Out[114]:
```
count    119.000000
mean     156.563025
std       11.370225
min      138.000000
25%      147.500000
50%      155.000000
75%      164.000000
max      185.000000
Name: bp_before, dtype: float64
```

In [115]:
```
df_wide["agegrp"].value_counts()
```

Out[115]:
```
46-59    40
60+      40
30-45    39
Name: agegrp, dtype: int64
```

In [116]:
```
df_wide["sex"].value_counts()
```

Out[116]:
```
Female    60
Male      59
Name: sex, dtype: int64
```

In [117]:
```
df_wide.head()
```

Out[117]:

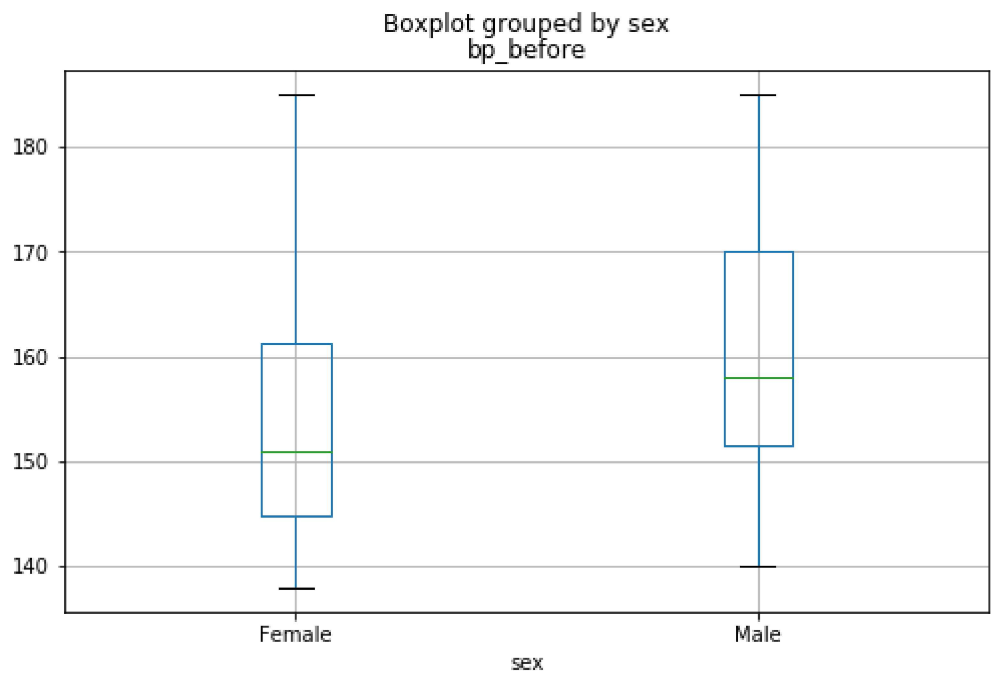
	patient	bp_after	bp_before	sex	agegrp
1	2	170	163	Male	30-45
2	3	168	153	Male	30-45
3	4	142	153	Male	30-45
4	5	141	146	Male	30-45
5	6	147	150	Male	30-45

Hypothesis test for relationship between sex and blood pressure based on before experiment blood pressure

In this section I will conduct hypothesis test on wheather male and female have the same or different blood pressure values.I will follow the following steps (As highlighted earlier I will use the before bp measure to avoid confounding with the experiment):

- Descriptive statistics and visualization of blood pressure between male and female to see if there is a difference
- Hypothesis test will use t-statistics with the null hypothesis of male and female have same blood pressure and alternative of different bp measures.

In [118]:
```
# box plots
df_wide.boxplot(column="bp_before",by="sex", figsize=(8,5));
```



In [119]:
```
df_wide.groupby('sex')['bp_before'].describe()
```

Out[119]:

sex	count	mean	std	min	25%	50%	75%	max
Female	60.0	153.633333	10.735600	138.0	144.75	151.0	161.25	185.0
Male	59.0	159.542373	11.308102	140.0	151.50	158.0	170.00	185.0

As can be seen from the plot and the descriptive statistics female tend to have lower mean, median and standard deviation of blood pressure compared to male. Next I will test whether the difference is significant by using t-test by checking the following:

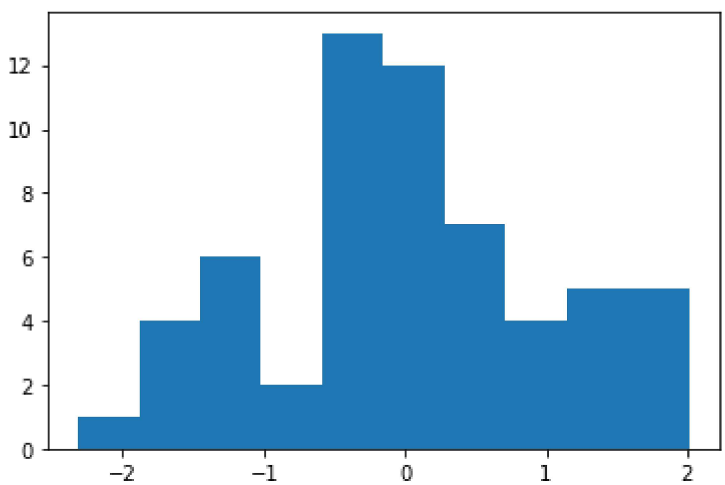
- Equal sample size
- Normality of the blood pressure difference series
- Equal variance of blood pressure between male and female

In [120]:
```
# (1): Equal sample size: as shown above in the descriptive statistics whe hav
e 60 female and 59 male - we randomly sample from
# the female 59 (drop 1)
male = np.array(df_wide[df_wide["sex"]=="Male"]['bp_before'])
female = np.array(df_wide[df_wide["sex"]=="Female"]['bp_before'].sample(59))
# blood pressure variable will be rescaled to standard normal
diff = scale(male - female)
```

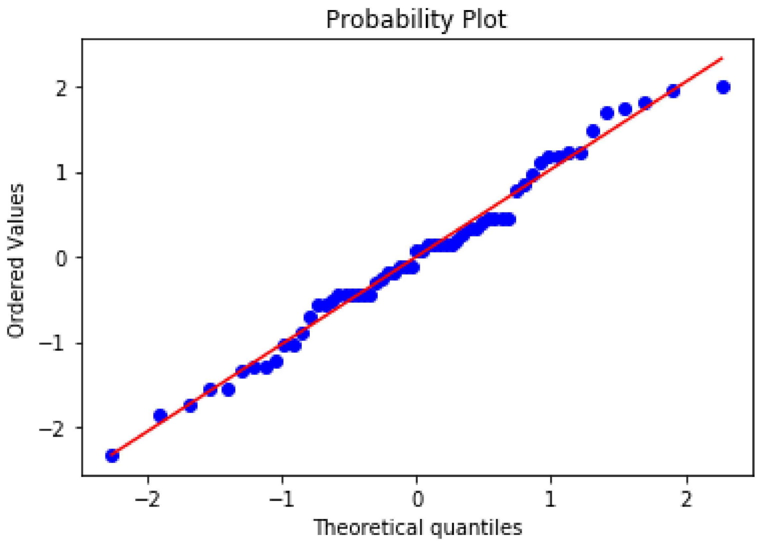
In [121]:
```
(len(male) == len(female) == len(diff))
```

Out[121]:
```
True
```

In [122]:
```
# (2) normality of bllod pressure difference series will be checked by histogr
am, qq plots and shapiro test
plt.hist(diff);
```



In [123]:
```
stats.probplot(diff, plot=plt, dist='norm');
```



In [124]:
```
stats.shapiro(diff)
```

Out[124]:
```
(0.9836707711219788, 0.6124415397644043)
```

In [125]:
```
# shapiro normality results show that the differences can be assumed to follow normal distribution at 10%  significance
# (3) check for equality of variances
stats.levene(male, female)
```

Out[125]:
```
LeveneResult(statistic=0.24486445282723712, pvalue=0.6216505702277931)
```

Leven's test for equality of variances show that the variances are equal with a large p-value of 0.55. Since now we have the assumptions for t-test (equal sample size, equal variance, normality of differences) we will apply a formal t-test to see if the results shown via descriptive statistics are not as a result of chance

In [126]:
```
stats.ttest_ind(male, female)
```

Out[126]:
```
Ttest_indResult(statistic=2.88576779009842, pvalue=0.004657315042889608)
```

the t-test statistic is significant with a p-value of 0.006. The resutls above show that **means of blood pressure (before) varies between male and female with men having higher blood pressure values**

Hypothesis test for relationship between age group and blood pressure based on before experiment blood pressure

In this section I will conduct hypothesis test on wheather the three age groups have the same or different blood pressure values.I will follow the following steps (As highlighted earlier I will use the before bp measure to avoid confounding with the experiment):

- Descriptive statistics and visualization of blood pressure between the age groups to see if there is a difference
- **Apply oneway-anova to see if the groups have different blood pressures**

In [127]:
```
df_wide["agegrp"].value_counts()
```

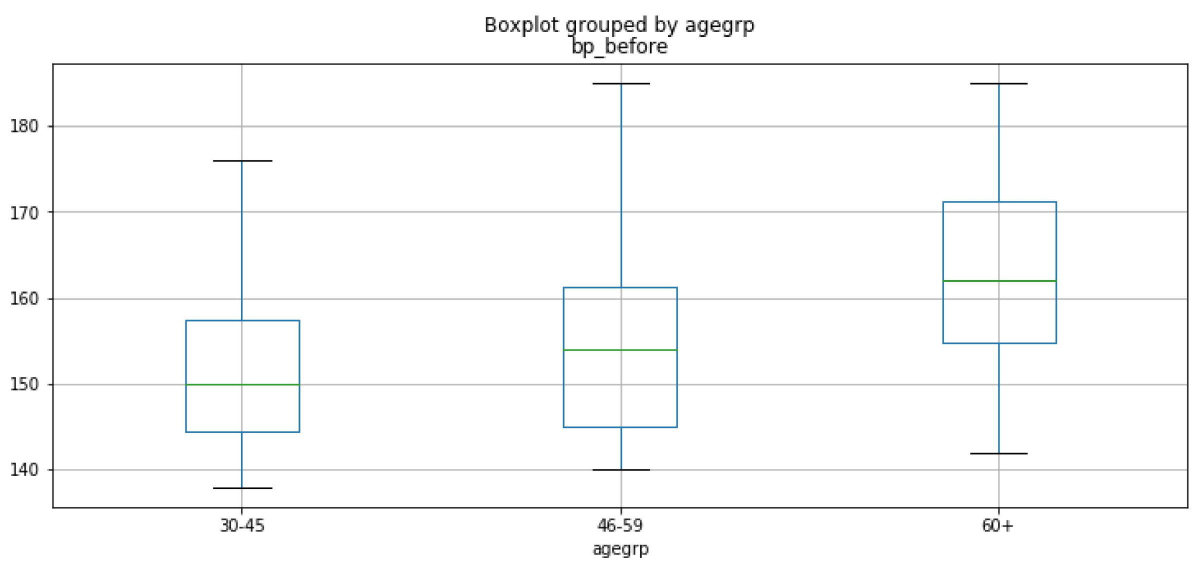
Out[127]:
```
46-59    40
60+      40
30-45    39
Name: agegrp, dtype: int64
```

In [128]:
```
df_wide.groupby("agegrp")["bp_before"].describe()
```

Out[128]:

agegrp	count	mean	std	min	25%	50%	75%	max
30-45	39.0	151.897436	9.270198	138.0	144.50	150.0	157.50	176.0
46-59	40.0	155.100000	11.459628	140.0	145.00	154.0	161.25	185.0
60+	40.0	162.575000	10.727122	142.0	154.75	162.0	171.25	185.0

In [129]:
```
df_wide.boxplot(column="bp_before", by="agegrp", figsize=(12,5));
```



The descriptive statistics and box plot show that as age increases blood pressure mean tends to increase. I will apply one-way anova to see if the groups differ below

In [130]:
```
# take equal sampe size from groups
age_3045 = df_wide[df_wide["agegrp"]=="30-45"]['bp_before']
age_4659 = df_wide[df_wide["agegrp"]=="46-59"]['bp_before'].sample(39)
age_60 = df_wide[df_wide["agegrp"]=="60+"]['bp_before'].sample(39)
```

In [131]:
```
stats.f_oneway(age_3045, age_4659, age_60)
```

Out[131]:
```
F_onewayResult(statistic=10.041325239652599, pvalue=9.623634218967992e-05)
```

The test statistics of the oneway-anova shows that the groups have different blood pressure values with a p-value of <0.01. **Multiple comparison test conducted between the groups show that there is no significant difference on blood pressure between age groups 30-45 and 46-59 while age group 60+ has different blood pressure compared to the two groups (30-45 and 46-59)**

In [132]:
```
# one way anova doesn't tell us how the means compare
from statsmodels.stats.multicomp import MultiComparison
mul_com = MultiComparison(df_wide['bp_before'], df_wide['agegrp'])
mul_result = mul_com.tukeyhsd()
print(mul_result)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
 30-45  46-59   3.2026 0.3714 -2.4265  8.8317  False
 30-45    60+  10.6776  0.001  5.0485 16.3067   True
 46-59    60+    7.475 0.0054  1.8816 13.0684   True
----------------------------------------------------
```

Two-way anova for relationship of sex and age-group with blood pressure

In this section, I applied OLS regression by modelling blood pressure (before) as dependent variable and sex and age group (including interations) as independent variables. The results below show that age group and sex affect blood pressure levels with the significance F-statistics. The lower adjusted R-squared value implies that there are also other factors that explain blood pressure. Sex and age group are individually significant factors but their interaction is insignificant.

In [133]:
```
model = ols('bp_before ~C(sex)*C(agegrp)', df_wide).fit()
print(model.summary())
```

```
                            OLS Regression Results                            
==============================================================================
=
Dep. Variable:              bp_before   R-squared:                       0.22
7
Model:                            OLS   Adj. R-squared:                  0.19
3
Method:                 Least Squares   F-statistic:                     6.64
2
Date:                Tue, 19 Jan 2021   Prob (F-statistic):          1.86e-0
5
Time:                        21:33:06   Log-Likelihood:                -442.3
1
No. Observations:                 119   AIC:                             896.
6
Df Residuals:                     113   BIC:                             913.
3
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
==============================================================================
========================
                                     coef    std err          t      P>|t|   
[0.025      0.975]
------------------------------------------------------------------------------
------------------------
Intercept                        149.9000      2.284     65.629      0.000   
145.375     154.425
C(sex)[T.Male]                     4.1000      3.272      1.253      0.213   
-2.383      10.583
C(agegrp)[T.46-59]                 1.2500      3.230      0.387      0.699   
-5.149       7.649
C(agegrp)[T.60+]                   9.9500      3.230      3.080      0.003   
3.551      16.349
C(sex)[T.Male]:C(agegrp)[T.46-59]     3.8000      4.598      0.826      0.410   
-5.310      12.910
C(sex)[T.Male]:C(agegrp)[T.60+]     1.3500      4.598      0.294      0.770   
-7.760      10.460
==============================================================================
=
Omnibus:                        6.384   Durbin-Watson:                   2.07
4
Prob(Omnibus):                  0.041   Jarque-Bera (JB):                6.59
0
Skew:                           0.569   Prob(JB):                       0.037
1
Kurtosis:                       2.820   Cond. No.                         9.7
9
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [134]:
```
sm.stats.anova_lm(model)
```

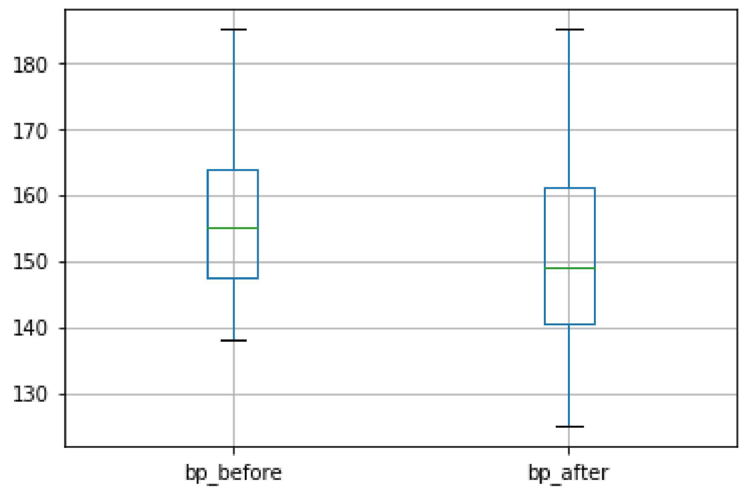
Out[134]:

	df	sum_sq	mean_sq	F	PR(>F)
C(sex)	1.0	1038.699910	1038.699910	9.955267	0.002056
C(agegrp)	2.0	2353.072875	1176.536438	11.276340	0.000034
C(sex):C(agegrp)	2.0	73.454526	36.727263	0.352007	0.704044
Residual	113.0	11790.050000	104.336726	NaN	NaN

Paired-ttest for difference in before and after experiment blood pressures

As the box plot and descriptive statistics beow show the blood pressure tends to reduce after the experiment

In [135]:
```
df_wide[['bp_before','bp_after']].boxplot();
```



In [136]:
```
df_wide[['bp_before','bp_after']].describe().T
```

Out[136]:

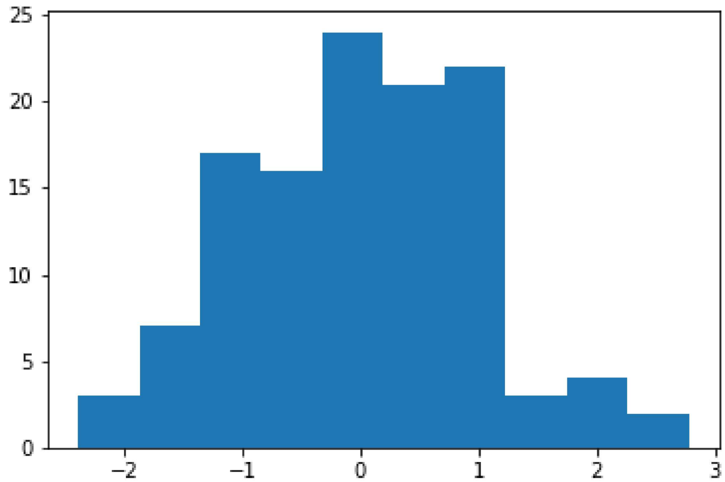
	count	mean	std	min	25%	50%	75%	max
bp_before	119.0	156.563025	11.370225	138.0	147.5	155.0	164.0	185.0
bp_after	119.0	151.344538	14.236761	125.0	140.5	149.0	161.0	185.0

In [137]:
```
# paired-ttest after checking equality of variances: variances are not equal
stats.levene(df_wide['bp_before'], df_wide['bp_after'])
```

Out[137]: LeveneResult(statistic=5.174239097139438, pvalue=0.023822348921505122)

In [138]:
```
diff = scale(df_wide['bp_after'] - df_wide['bp_before'])
print("Histogram of blood pressure difference (after-before)")
plt.hist(diff);
```

```
Histogram of blood pressure difference (after-before)
```



In [139]:
```
stats.shapiro(diff) # differences are normally differences
```

Out[139]:
```
(0.9927873015403748, 0.7977616786956787)
```

In [140]:
```
stats.ttest_rel(df_wide['bp_before'], df_wide['bp_after'])
```

Out[140]:
```
Ttest_relResult(statistic=3.4034635549438716, pvalue=0.0009094876218154547)
```

The test statistics and p-value above show that the experiment indeed reduced the blood pressure.

Conclusion

In this project I have used Stata statistical software fictitious blood pressure data to examine the relationship between age and sex on blood pressure and the impact of the experiment on blood pressure levels. Here are the results from the analysis:

- **Males tend to have higher blood pressure (bp) values compared to females:** the results from descriptive statistics and statistical test show that males have higher blood pressure compared to females. This requires further study on why men are more succeptible.
- **One-way anova resutls show that older individuals tend to have higher bp:** Comparison across the age groups showed that 60+ age group has higher age group compared to age groups 30-45 and 46-59 while there is no statistically significant difference between age groups 30-45 and 46-59
- **Two-way anova results show that both sex and age group affect blood pressure while their interaction is insignificant**

Next steps for data analysis:

It is important to understand the following points in-depth in the future:

- what is the reason for men to have higher bp compared to female
- Conducting the analysis again by categorizing the age groups again as now there is no difference between 30-45 and 46-59 but having access to detailed age data and exploring the distributions of the age groups is good to confirm whether they are indeed similar. Probably the result might be because the groups are not representative or sampe size is small
- Exploring other variables that affect blood pressure: the OLS results show that the adjusted R-squared is small(0.19) which implies there are other variables that explain blood pressure in addition to the two.

-- Thanks for reading and giving your feedback! ----------------------------------