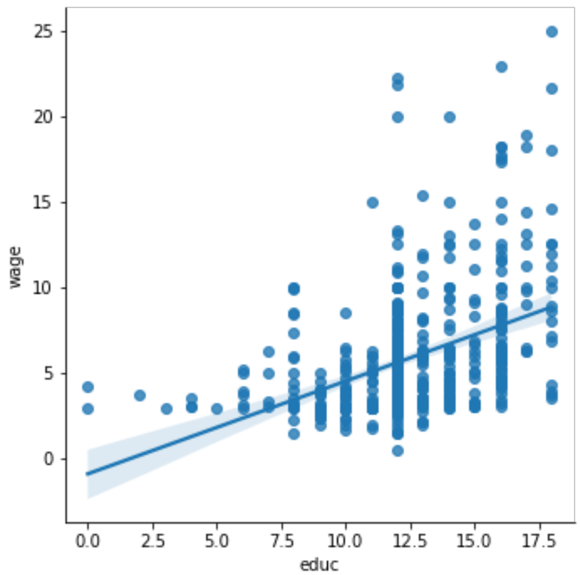
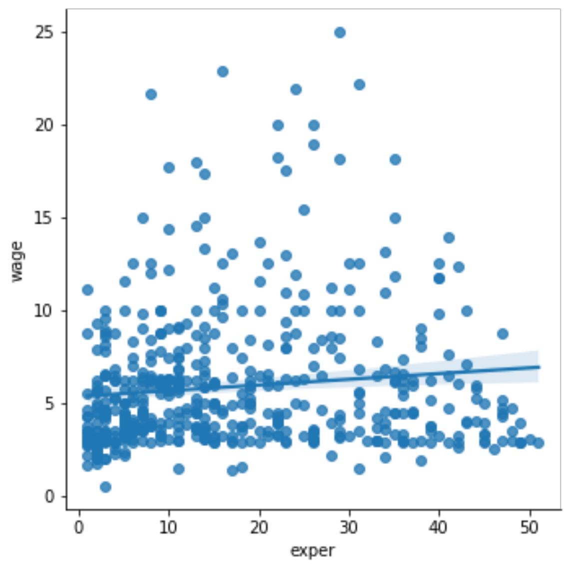
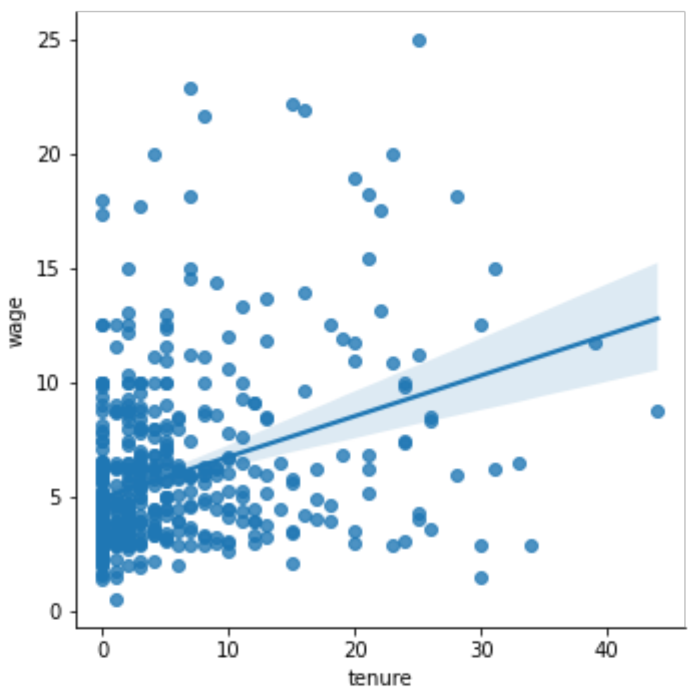
Problem Set 5

Exercise 1

- Part 2







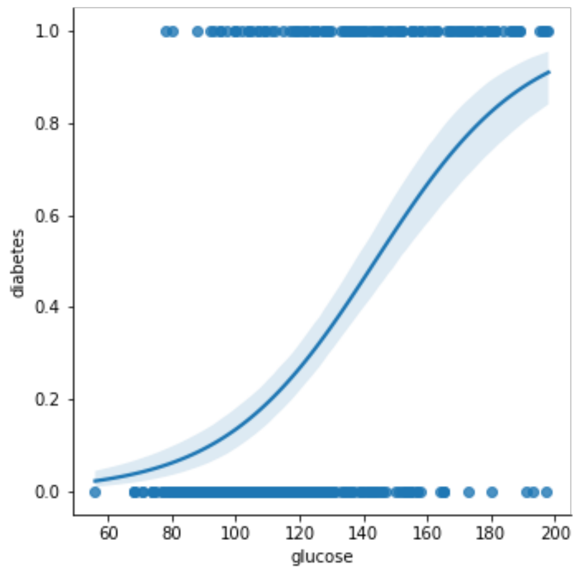
- Part 3
    - I think using OLS Regression is more suitable in understanding the factors that explain variability in wages because wage/logged wage is the response variable (LHS) of a regression model, and it is continuous. Logit regression is also used when the response variable has multiple outcomes, which isn't really applicable here.
- Part 4
    - Since we are using wages as the dependent variable on the left-hand side and we are trying to find the correlation between wages and other variables, I've created a model that checks the relationship between wage and all the continuous variables in the wage dataset, as well as two binary variables. I put those variables in an array and used the sm approach.
    - LHS: wage; RHS: education, experience, tenure, nonwhite, female
- Part 5

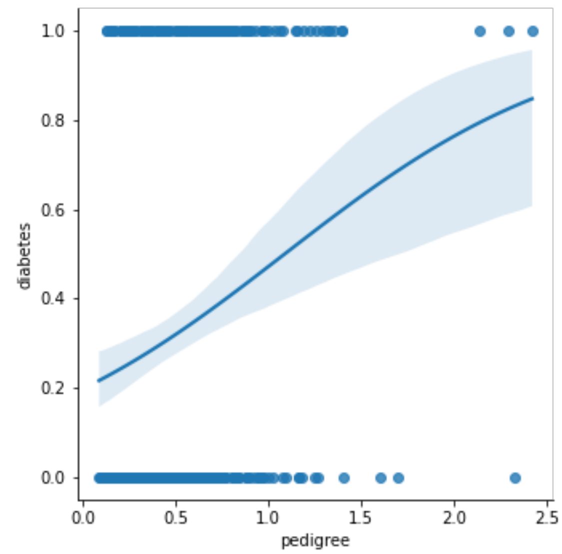
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.364
Model:                            OLS   Adj. R-squared:                  0.358
Method:                 Least Squares   F-statistic:                     59.43
Date:                Mon, 22 Nov 2021   Prob (F-statistic):           6.48e-49
Time:                        23:47:40   Log-Likelihood:                -1314.2
No. Observations:                 526   AIC:                             2640.
Df Residuals:                     520   BIC:                             2666.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.5403      0.732     -2.103      0.036      -2.979      -0.102
educ           0.5703      0.050     11.507      0.000       0.473       0.668
exper          0.0253      0.012      2.188      0.029       0.003       0.048
tenure         0.1411      0.021      6.660      0.000       0.099       0.183
nonwhite      -0.1159      0.427     -0.271      0.786      -0.955       0.723
female        -1.8120      0.265     -6.835      0.000      -2.333      -1.291
==============================================================================
Omnibus:                      185.640   Durbin-Watson:                   1.794
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              712.187
Skew:                           1.588   Prob(JB):                    2.24e-155
Kurtosis:                       7.733   Cond. No.                         143.
==============================================================================
```

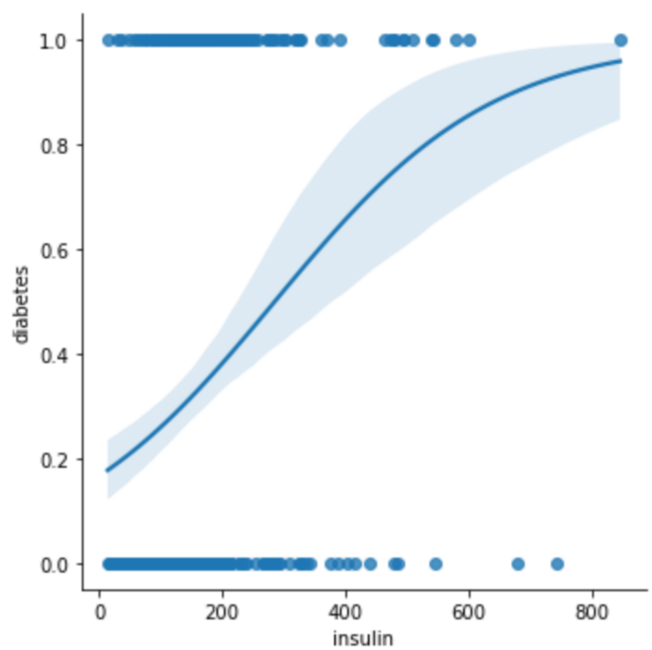
- Part 6
    - Education, experience, and tenure all have positive correlations with wages, and they're all statistically significant because their p-values are all lower than 0.05. Variable female has a negative correlation with wage and it's also statistically significant because of its low p-value. The only nonsignificant variable is nonwhite.
- Part 7
    - Since the OLS regression yields an R-squared value of 0.364, this means that 36.4% of variation in the wage dataset can be explained by the variables educ, exper, tenure, nonwhite and female.
- Part 8
    - "wage": [150], "exper": [100], "tenure": [50], “educ": [100], "female": [0], "nonwhite": [0]})

Exercise 2

- Part 2







- Part 3
    - Logistic Regression is more suitable to analyze this problem because diabetes is a binary variable, meaning there are two outcomes, and each outcome has a probability between 0 and 1.
- Part 4
    - Since we are using diabetes as the dependent variable on the left-hand side and we are trying to find the correlation between diabetes and other variables, I've created a model that checks the relationship between diabetes and all the variables in the diabetes dataset. I put those variables in an array and used the sm approach.
    - LHS: diabetes, RHS: pregnant, glucose, pressure, triceps, insulin, mass, pedigree, age
- Part 5

```
Optimization terminated successfully.
         Current function value: 0.438803
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:               diabetes   No. Observations:                  392
Model:                          Logit   Df Residuals:                      383
Method:                           MLE   Df Model:                            8
Date:                Tue, 23 Nov 2021   Pseudo R-squ.:                  0.3093
Time:                        00:06:15   Log-Likelihood:                -172.01
converged:                       True   LL-Null:                       -249.05
Covariance Type:            nonrobust   LLR p-value:                 2.765e-29
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -10.0407      1.218     -8.246      0.000     -12.427      -7.654
pregnant       0.0822      0.055      1.482      0.138      -0.026       0.191
glucose        0.0383      0.006      6.635      0.000       0.027       0.050
pressure      -0.0014      0.012     -0.120      0.904      -0.025       0.022
triceps        0.0112      0.017      0.657      0.511      -0.022       0.045
insulin       -0.0008      0.001     -0.632      0.528      -0.003       0.002
mass           0.0705      0.027      2.580      0.010       0.017       0.124
pedigree       1.1409      0.427      2.669      0.008       0.303       1.979
age            0.0340      0.018      1.847      0.065      -0.002       0.070
==============================================================================
```

- Part 6
    - Glucose, mass and pedigree all have positive correlations with diabetes, and they're all statistically significant because their p-values are all lower than 0.05. The rest of the variables aren't statistically significant because their p-values are higher than 0.05.
- Part 7
    - The patient in the median of each of my independent variables has a 19.3% chance of having diabetes while the patient in the 75% tile has a 64.9% chance and the patient in the 25% tile has a 4.78% chance.

Code

```python
"""This is the code file for both portions of problem set 5."""
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import statsmodels.api as sm


# Exercise 1
def first_exercise():
    """
    This function loads and preps the dataset that contains information about
    expected wage rates and its related variables. It also generates visuals
    that describe the relationship between wage and years of education. It
    also generates a model for wage data and estimates such model by producing
    an OLS regression table.
    """
    # Part 1
    wage = pd.read_csv("wage.csv").dropna()
    # Part 2
    sns.lmplot(x='educ', y='wage', data=wage)
    fig1 = plt.figure(1)
    fig1.savefig("1_2_1.png")
    sns.lmplot(x='exper', y='wage', data=wage)
    fig2 = plt.figure(2)
    fig2.savefig("1_2_2.png")
    sns.lmplot(x='tenure', y='wage', data=wage)
    fig3 = plt.figure(3)
    fig3.savefig("1_2_3.png")
    # Part 5
    lhs = wage['wage']
    ind_vars = ['educ', 'exper', 'tenure', 'nonwhite', 'female']
    rhs = wage.loc[:, ind_vars]
    rhs = sm.add_constant(rhs)
    mod = sm.OLS(lhs, rhs)
    res = mod.fit()
    print(res.summary())
    # Part 8
    hypo = pd.DataFrame({"wage": [150], "exper": [100], "tenure": [50],
                    "educ": [100], "female": [0], "nonwhite": [0]})
    print(res.predict(hypo))
```

```
# Exercise 2
def second_exercise():
    """
    This function loads and preps the dataset that contains information about
    diabetes rates and its related variables. It also generates visuals
    that describe the relationship between diabetes and variables like insulin
    levels, pedigree and glucose. It also generates a model for diabetes data
    and estimates such model by producing a logistic regression table.
    """
    # Part 1
    diabetes = pd.read_csv("diabetes.csv").dropna()
    diabetes['diabetes'] = diabetes.diabetes.map({'pos': 1, 'neg': 0})
    # Part 2
    sns.lmplot(x="insulin", y="diabetes", data=diabetes, logistic=True)
    fig4 = plt.figure(4)
    fig4.savefig("2_2_1.png")
    sns.lmplot(x="pedigree", y="diabetes", data=diabetes, logistic=True)
    fig5 = plt.figure(5)
    fig5.savefig("2_2_2.png")
    sns.lmplot(x="glucose", y="diabetes", data=diabetes, logistic=True)
    fig6 = plt.figure(6)
    fig6.savefig("2_2_3.png")
    # Part 5
    lhs = diabetes['diabetes']
    ind_vars = ['pregnant', 'glucose', 'pressure', 'triceps',
                'insulin', 'mass', 'pedigree', 'age']
    rhs = diabetes.loc[:, ind_vars]
    rhs = sm.add_constant(rhs)
    mod = sm.Logit(lhs, rhs)
    res = mod.fit()
    print(res.summary())
    # Part 7
    for percentiles in [.25, .5, .75]:
        values = rhs.quantile(percentiles)
        print(res.predict(values.to_numpy()))
```