# coding_challenge

### CSE 250 Coding Challenge

Ammon Van Engelenhoven

### Challenge Summary

After completing the challenge, describe how you think you did.

#### Challenge 1

The below code reads in the data from one of our projects this semester. In addition, it removes all rows that don't report a month. Use another Pandas method(s) besides .replace() and dropna() to arrive at the same data set.

##### Answer

Include the image, answer, or table here

##### Code

```
def gq_1_example():
    base_url = 'https://github.com/byuidatascience/data4missing/'
    flights_path = 'raw/master/data-raw/flights_missing/flights_missing.json'
    url_flights = base_url + flights_path
    flights = (pd.read_json(url_flights)
        .replace({"month":{"n/a":np.nan}})
        .dropna(subset=['month']))
    print(flights.info())

gq_1_example()

def gq_1_answer():
    base_url = 'https://github.com/byuidatascience/data4missing/'
    flights_path = 'raw/master/data-raw/flights_missing/flights_missing.json'
    url_flights = base_url + flights_path

    df = pd.read_json(url_flights)

    df = df[df["month"] != "n/a"]
    print(df.info())
    return

gq_1_answer()
```

#### Challenge 2

Two of the minutes columns have missing values. Identify those two columns and replace the missing values with the respective median of those columns. After you have fixed the missing values in those two columns report the standard deviation of each column.

##### Answer

This was rather simple to solve after pulling the Data in from Question 1 I noticed that 2 columns were not the same as the rest for the minute data.

So I filled in the NA with the mean of the column and set each of them in place of the np.nan.

##### Code

```
def GQ_2():
    base_url = 'https://github.com/byuidatascience/data4missing/'
    flights_path = 'raw/master/data-raw/flights_missing/flights_missing.json'
    url_flights = base_url + flights_path

    df = (pd.read_json(url_flights)
        .replace({"month":{"n/a":np.nan}})
        .dropna(subset=['month']))
    ## 2 columns are minutes_delayed_nas & minutes_delayed_carrier, because they have lower numbers then the rest of the "Minutes" data.

    df["minutes_delayed_nas"].fillna(float(df['minutes_delayed_nas'].mean()), inplace=True)
    df["minutes_delayed_carrier"].fillna(float(df['minutes_delayed_carrier'].mean()), inplace=True)

    df = df[["minutes_delayed_nas", "minutes_delayed_carrier"]]
    print(df.info())
    print(df.head())


    # Now Find the Standard Deviation of these 2 columns
    print("\n \n Standard Deviation of Minutes delayed NAS & Carrier: \n", round(df.std(), 2))


    return df


GQ_2()
```

#### Challenge 3

We want to convert month to a numeric value in our flights data to use in our machine learning model. Please update the month column (you can use month_dict if you would like) and report the mean of the month column.

##### Answer

I used the Data Sictionary to make it easier for my self when I passed it through with the Replace function.

##### Code

```
def GQ_3():
    base_url = 'https://github.com/byuidatascience/data4missing/'
    flights_path = 'raw/master/data-raw/flights_missing/flights_missing.json'
    url_flights = base_url + flights_path
    df = pd.read_json(url_flights)

    month_dict = {"January":1, "Febuary":2, "March":3, "April":4,
    "May":5, "June":6, "July":7, "August":8, "September":9,
    "October":10, "November":11, "December":12}

    df = df[df["month"] != "n/a"]

    df = df["month"].replace(month_dict, method='bfill').astype("int64")

    print("Median Month in the Data Set is the ", round(df.mean(), 0), "th")
    return df

GQ_3()
```

#### Challenge 4

The num_of_delays_carrier column has a Dtype of object. We need to remove the + that is causing the dtype and convert it to an float64. Fix that issue in the flights data. Report the mean of num_of_delays_carrier.

##### Answer

Did not finish the question. Was trying to use regex to solve the problem but I had a hard time remembering what symbols to use for the Regex for the "+" sign.

##### Code

```
def GQ_4():
    base_url = 'https://github.com/byuidatascience/data4missing/'
    flights_path = 'raw/master/data-raw/flights_missing/flights_missing.json'
    url_flights = base_url + flights_path

    df = pd.read_json(url_flights)

    df["num_of_delays_carrier"] =  df["num_of_delays_carrier"].map(lambda x: x.lstrip('+-').rstrip('aAbBcC'))
    df["num_of_delays_carrier"].astype("float64")
    print(df.info())
    return df

GQ_4()
```