final_project

September 29, 2020

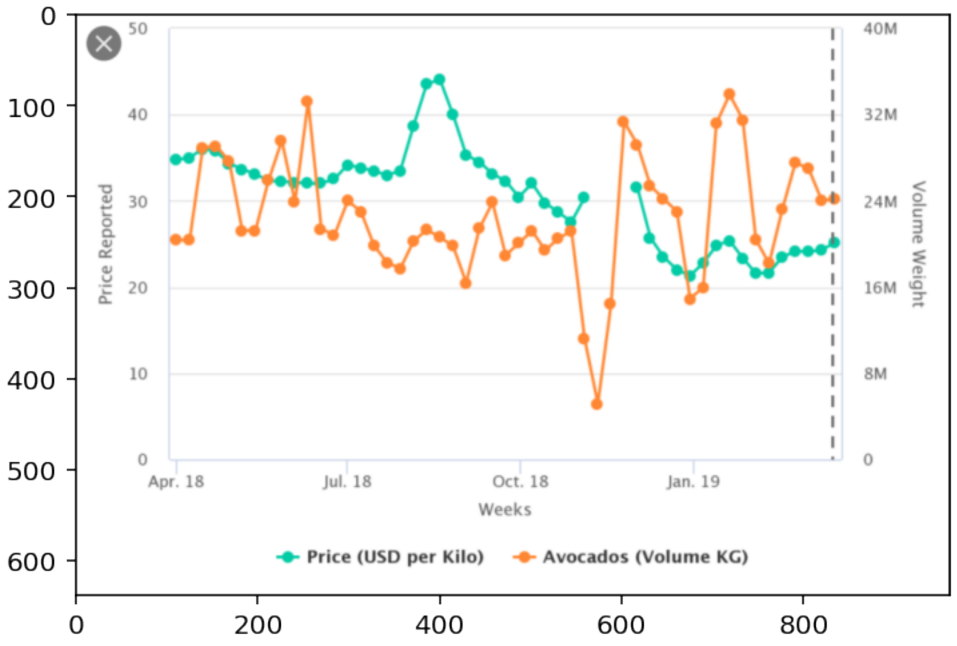
[21]:

```
from datetime import datetime
import numpy as np
import pandas as pd
import matplotlib
import matplotlib as mpl
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
import matplotlib.patches as mpatches
import matplotlib.patches as patches # for drawing shapes


%config InlineBackend.figure_format = 'retina'
```

1 1. Avocado Price

[22]:

```
# Load the orignal plot
img = mpimg.imread('data/bad_plot.png')
plt.imshow(img)
plt.show()
```



[23]:

```
# Data source and columns explanation
# https://www.kaggle.com/neuromusic/avocado-prices
```

[24]:

```
# Data from https://www.kaggle.com/neuromusic/avocado-prices
df = pd.read_csv("data/avocado.csv")

# Convert date column to datetime object, for plotting purposes
df['Date']= pd.to_datetime(df['Date'])

# Normalize the total volume column, for plotting purposes
df['Toal_Volume_normalized']= df["Total Volume"]/1000
df.head()
```

[24]:

```
   Unnamed: 0       Date  AveragePrice  Total Volume     4046       4225  \
0           0 2015-12-27          1.33      64236.62  1036.74   54454.85
1           1 2015-12-20          1.35      54876.98   674.28   44638.81
2           2 2015-12-13          0.93     118220.22   794.70  109149.67
3           3 2015-12-06          1.08      78992.15  1132.00   71976.41
4           4 2015-11-29          1.28      51039.60   941.48   43838.39

     4770  Total Bags  Small Bags  Large Bags  XLarge Bags          type  \
0   48.16     8696.87     8603.62       93.25          0.0  conventional
1   58.33     9505.56     9408.07       97.49          0.0  conventional
2  130.50     8145.35     8042.21      103.14          0.0  conventional
3   72.58     5811.16     5677.40      133.76          0.0  conventional
```

```
4   75.78     6183.95     5986.26      197.69           0.0  conventional

   year  region  Toal_Volume_normalized
0  2015  Albany                64.23662
1  2015  Albany                54.87698
2  2015  Albany               118.22022
3  2015  Albany                78.99215
4  2015  Albany                51.03960
```

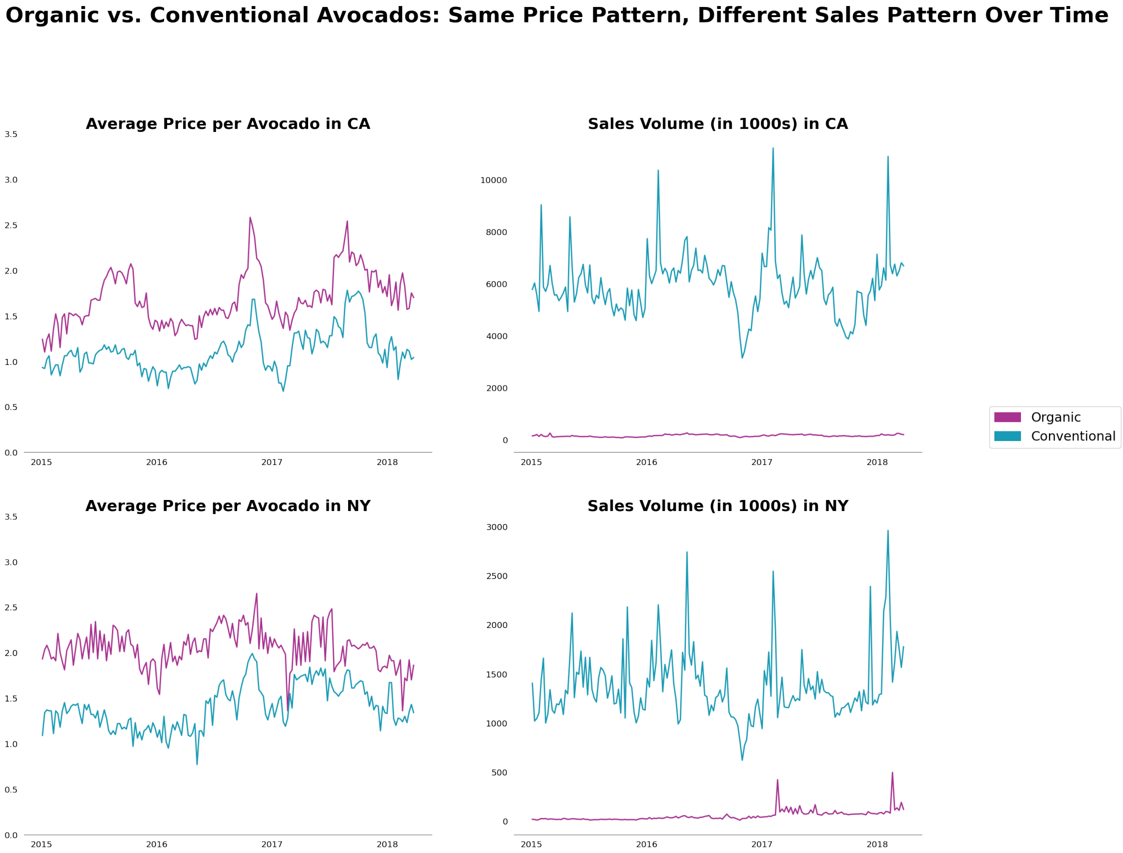
[25]:
```
# Focus on California and New York data only
df_CA = df[df["region"]=="California"]
df_NY = df[df["region"]=="NewYork"]
```

[26]:
```
# Sort the dataframe by date for CA and NY
df_CA = df_CA.sort_values(by='Date')
df_NY = df_NY.sort_values(by='Date')

# Data on convetional vs. organic avocado in CA
df_CA_conventional = df_CA[df_CA["type"] == "conventional"]
df_CA_organic = df_CA[df_CA["type"] == "organic"]

# Data on convetional vs. organic avocado in NY
df_NY_conventional = df_NY[df_NY["type"] == "conventional"]
df_NY_organic = df_NY[df_NY["type"] == "organic"]
```

[27]:
```
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(20,15))
axes = axes.flatten()



# Plot california AVERAGE RICE TREND OVER TIME
# Set the y axis limit within the subplot to smooth the trend line
axes[0].set_ylim([0, 3.5])
axes[0].plot(df_CA_organic["Date"],df_CA_organic["AveragePrice"],c="#a83290")
axes[0].
 ↪plot(df_CA_conventional["Date"],df_CA_conventional["AveragePrice"],c="#199bb5")
# Set the xlabel date locators
axes[0].xaxis.set_major_locator(matplotlib.dates.YearLocator())
axes[0].xaxis.set_major_formatter(matplotlib.dates.DateFormatter('%Y'))
axes[0].set_title("Average Price per Avocado in CA",fontweight="bold",size=18)



# Plot california TOTAL VOLUME TREND OVER TIME
axes[1].
 ↪plot(df_CA_organic["Date"],df_CA_organic["Toal_Volume_normalized"],c="#a83290")
axes[1].
 ↪plot(df_CA_conventional["Date"],df_CA_conventional["Toal_Volume_normalized"],c="#199bb5")
# Set the xlabel date locators
```

```
axes[1].xaxis.set_major_locator(matplotlib.dates.YearLocator())
axes[1].xaxis.set_major_formatter(matplotlib.dates.DateFormatter('%Y'))
axes[1].set_title("Sales Volume (in 1000s) in CA",fontweight="bold",size=18)

# Plot NY AVERAGE PRICE TREND OVER TIME
axes[2].set_ylim([0, 3.5])
axes[2].plot(df_NY_organic["Date"],df_NY_organic["AveragePrice"],c="#a83290")
axes[2].
 ↪plot(df_NY_conventional["Date"],df_NY_conventional["AveragePrice"],c="#199bb5")
# Set the xlabel date locators
axes[2].xaxis.set_major_locator(matplotlib.dates.YearLocator())
axes[2].xaxis.set_major_formatter(matplotlib.dates.DateFormatter('%Y'))
axes[2].set_title("Average Price per Avocado in NY",fontweight="bold",size=18)

# Plot NY TOTAL VOLUME TREND OVER TIME
axes[3].
 ↪plot(df_NY_organic["Date"],df_NY_organic["Toal_Volume_normalized"],c="#a83290")
axes[3].
 ↪plot(df_NY_conventional["Date"],df_NY_conventional["Toal_Volume_normalized"],c="#199bb5")
# Set the xlabel date locators
axes[3].xaxis.set_major_locator(matplotlib.dates.YearLocator())
axes[3].xaxis.set_major_formatter(matplotlib.dates.DateFormatter('%Y'))
axes[3].set_title("Sales Volume (in 1000s) in NY",fontweight="bold",size=18)

# Set the format for the axis for all subplots
for i in range(4):
    axes[i].spines['left'].set_visible(False)
    axes[i].spines['top'].set_visible(False)
    axes[i].spines['right'].set_visible(False)
    axes[i].spines['bottom'].set_linewidth(.5)
    axes[i].yaxis.set_ticks_position('none')
    axes[i].xaxis.set_ticks_position('none')

# Add the descriptive title
fig.text(0.11,1,"Organic vs. Conventional Avocados: Same Price Pattern,␣
 ↪Different Sales Pattern Over Time ", fontweight='bold',fontsize=25)

# Add the legend
# https://stackoverflow.com/questions/9834452/
 ↪how-do-i-make-a-single-legend-for-many-subplots-with-matplotlib
labels = ['Organic', 'Conventional']
# now, create an artist for each color
organic_patch = mpatches.Patch(facecolor='#a83290', edgecolor='#a83290') #this␣
 ↪will create a red bar with black borders, you can leave out edgecolor if you␣
 ↪do not want the borders
conventional_patch = mpatches.Patch(facecolor='#199bb5', edgecolor='#199bb5')
```

```
fig.legend(handles = [organic_patch, conventional_patch,],labels=labels,
       loc="center right",
       borderaxespad=0.1,
        prop={'size': 15})
plt.subplots_adjust(right=0.85) #adjust the subplot to the right for the legend

plt.savefig("Avocado.png")
```


Organic vs. Conventional Avocados: Same Price Pattern, Different Sales Pattern Over Time

Average Price per Avocado in CA

Sales Volume (in 1000s) in CA

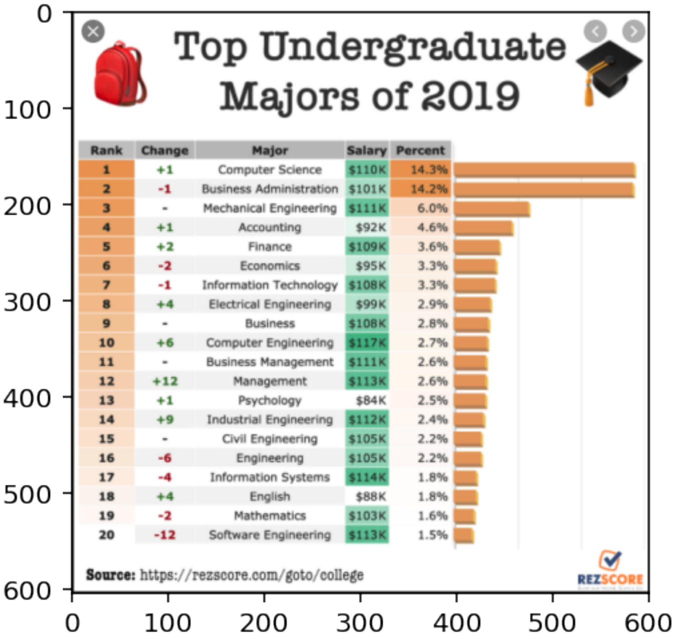
Average Price per Avocado in NY

Sales Volume (in 1000s) in NY

2 2. Salary vs. major/college/region

[28]:
```
# https://www.kaggle.com/wsj/college-salaries
# https://rstudio-pubs-static.s3.amazonaws.com/
 ↪343920_b2f5f1d787384dcaa97c4bcdb602a4ae.html
```

[29]:
```
# Orignal plot
img = mpimg.imread('data/bad_plot2.png')
plt.imshow(img)
plt.show()
```


0
100
200
300
400
500
600
0
100
200
300
400
500
600

Top Undergraduate Majors of 2019

Rank	Change	Major	Salary	Percent
1	+1	Computer Science	$110K	14.3%
2	-1	Business Administration	$101K	14.2%
3	-	Mechanical Engineering	$111K	6.0%
4	+1	Accounting	$92K	4.6%
5	+2	Finance	$109K	3.6%
6	-2	Economics	$95K	3.3%
7	-1	Information Technology	$108K	3.3%
8	+4	Electrical Engineering	$99K	2.9%
9	-	Business	$108K	2.8%
10	+6	Computer Engineering	$117K	2.7%
11	-	Business Management	$111K	2.6%
12	+12	Management	$113K	2.6%
13	+1	Psychology	$84K	2.5%
14	+9	Industrial Engineering	$112K	2.4%
15	-	Civil Engineering	$105K	2.2%
16	-6	Engineering	$105K	2.2%
17	-4	Information Systems	$114K	1.8%
18	+4	English	$88K	1.8%
19	-2	Mathematics	$103K	1.6%
20	-12	Software Engineering	$113K	1.5%

Source: https://rezscore.com/goto/college

```
[30]: salary_college_df = pd.read_csv("data/salaries-by-college-type.csv")
      salary_college_df.head()
```

```
[30]:                                         School Name  School Type  \
      0   Massachusetts Institute of Technology (MIT)  Engineering
      1      California Institute of Technology (CIT)  Engineering
      2                           Harvey Mudd College  Engineering
      3  Polytechnic University of New York, Brooklyn  Engineering
      4                                  Cooper Union  Engineering

        Starting Median Salary Mid-Career Median Salary  \
      0             $72,200.00              $126,000.00
      1             $75,500.00              $123,000.00
      2             $71,800.00              $122,000.00
      3             $62,400.00              $114,000.00
      4             $62,200.00              $114,000.00

        Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
      0                        $76,800.00                        $99,200.00
      1                               NaN                       $104,000.00
      2                               NaN                        $96,000.00
      3                        $66,800.00                        $94,300.00
      4                               NaN                        $80,200.00

        Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
```

```
0                       $168,000.00                       $220,000.00
1                       $161,000.00                               NaN
2                       $180,000.00                               NaN
3                       $143,000.00                       $190,000.00
4                       $142,000.00                               NaN
```

```
[31]: # Convert the target column to float for plotting purposes
salary_college_df["Starting Median Salary"]=salary_college_df["Starting Median␣
 ↪Salary"].str.replace("$","").str.replace(",","")
salary_college_df["Starting Median Salary"]=salary_college_df["Starting Median␣
 ↪Salary"].astype(float)
salary_college_df.head()
```

```
[31]:                                    School Name  School Type  \
0   Massachusetts Institute of Technology (MIT)  Engineering
1      California Institute of Technology (CIT)  Engineering
2                           Harvey Mudd College  Engineering
3  Polytechnic University of New York, Brooklyn  Engineering
4                                  Cooper Union  Engineering

   Starting Median Salary Mid-Career Median Salary  \
0                 72200.0              $126,000.00
1                 75500.0              $123,000.00
2                 71800.0              $122,000.00
3                 62400.0              $114,000.00
4                 62200.0              $114,000.00

  Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                        $76,800.00                        $99,200.00
1                               NaN                       $104,000.00
2                               NaN                        $96,000.00
3                        $66,800.00                        $94,300.00
4                               NaN                        $80,200.00

  Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
0                       $168,000.00                       $220,000.00
1                       $161,000.00                               NaN
2                       $180,000.00                               NaN
3                       $143,000.00                       $190,000.00
4                       $142,000.00                               NaN
```

```
[32]: grouped_df=salary_college_df.groupby("School Type").mean()
grouped_df.sort_values(by=['Starting Median Salary'])
```

```
[32]:               Starting Median Salary
School Type
State                   44126.285714
```

```
Party                  45715.000000
Liberal Arts           45746.808511
Engineering            59057.894737
Ivy League             60475.000000
```

[33]:

```
# Categeorize starting median salary data by Schol Type
ivy=salary_college_df[salary_college_df['School Type']=="Ivy League"]["Starting␣
 ↪Median Salary"].values/1000
engineering=salary_college_df[salary_college_df['School␣
 ↪Type']=="Engineering"]["Starting Median Salary"].values/1000
lib_arts=salary_college_df[salary_college_df['School Type']=="Liberal␣
 ↪Arts"]["Starting Median Salary"].values/1000
party=salary_college_df[salary_college_df['School Type']=="Party"]["Starting␣
 ↪Median Salary"].values/1000
state=salary_college_df[salary_college_df['School Type']=="State"]["Starting␣
 ↪Median Salary"].values/1000
```

[34]:

```
# Data cleaning for another dataframe
degree_df=pd.read_csv("data/degrees-that-pay-back.csv")
degree_df.head()
```

[34]:

```
     Undergraduate Major Starting Median Salary Mid-Career Median Salary  \
0             Accounting             $46,000.00               $77,100.00
1  Aerospace Engineering             $57,700.00              $101,000.00
2            Agriculture             $42,600.00               $71,900.00
3           Anthropology             $36,800.00               $61,500.00
4           Architecture             $41,600.00               $76,800.00

   Percent change from Starting to Mid-Career Salary  \
0                                               67.6
1                                               75.0
2                                               68.8
3                                               67.1
4                                               84.6

  Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                        $42,200.00                        $56,100.00
1                        $64,300.00                        $82,100.00
2                        $36,300.00                        $52,100.00
3                        $33,800.00                        $45,500.00
4                        $50,600.00                        $62,200.00

  Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
0                       $108,000.00                       $152,000.00
1                       $127,000.00                       $161,000.00
2                        $96,300.00                       $150,000.00
3                        $89,300.00                       $138,000.00
```

```
4                        $97,000.00                       $136,000.00
```

[35]:
```
# Convert the target column to float for plotting purposes
degree_df["Starting Median Salary"]=degree_df["Starting Median Salary"].str.
↪replace("$","").str.replace(",","")
degree_df["Starting Median Salary"]=degree_df["Starting Median Salary"].
↪astype(float)
degree_df.head()
```

[35]:
```
     Undergraduate Major  Starting Median Salary Mid-Career Median Salary  \
0             Accounting                 46000.0               $77,100.00
1  Aerospace Engineering                 57700.0              $101,000.00
2            Agriculture                 42600.0               $71,900.00
3           Anthropology                 36800.0               $61,500.00
4           Architecture                 41600.0               $76,800.00

   Percent change from Starting to Mid-Career Salary  \
0                                              67.6
1                                              75.0
2                                              68.8
3                                              67.1
4                                              84.6

  Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
0                        $42,200.00                        $56,100.00
1                        $64,300.00                        $82,100.00
2                        $36,300.00                        $52,100.00
3                        $33,800.00                        $45,500.00
4                        $50,600.00                        $62,200.00

  Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
0                       $108,000.00                       $152,000.00
1                       $127,000.00                       $161,000.00
2                        $96,300.00                       $150,000.00
3                        $89,300.00                       $138,000.00
4                        $97,000.00                       $136,000.00
```

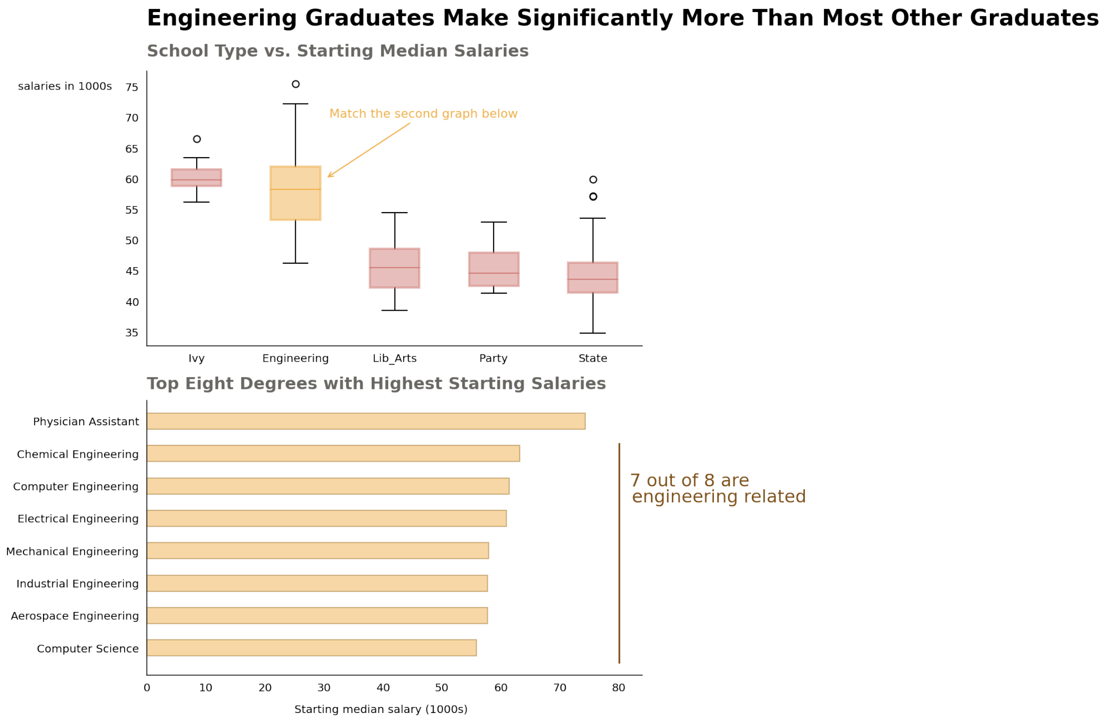
[36]:
```
# Extract the top ten high paying majors
top_ten_pay_df=degree_df.sort_values(by="Starting Median␣
↪Salary",ascending=False)[:8]
top_ten_pay_df["Starting Median Salary"]=top_ten_pay_df["Starting Median␣
↪Salary"]/1000.
top_ten_pay_df=top_ten_pay_df.sort_values("Starting Median␣
↪Salary",ascending=True)
top_ten_pay_df
```

[36]:
```
       Undergraduate Major  Starting Median Salary Mid-Career Median Salary  \
13        Computer Science                    55.9               $95,500.00
1      Aerospace Engineering                  57.7              $101,000.00
30    Industrial Engineering                  57.7               $94,700.00
38    Mechanical Engineering                  57.9               $93,600.00
19    Electrical Engineering                  60.9              $103,000.00
12      Computer Engineering                  61.4              $105,000.00
8       Chemical Engineering                  63.2              $107,000.00
43       Physician Assistant                  74.3               $91,700.00

    Percent change from Starting to Mid-Career Salary  \
13                                               70.8
1                                                75.0
30                                               64.1
38                                               61.7
19                                               69.1
12                                               71.0
8                                                69.3
43                                               23.4

   Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary  \
13                        $56,000.00                        $74,900.00
1                         $64,300.00                        $82,100.00
30                        $57,100.00                        $72,300.00
38                        $63,700.00                        $76,200.00
19                        $69,300.00                        $83,800.00
12                        $66,100.00                        $84,100.00
8                         $71,900.00                        $87,300.00
43                        $66,400.00                        $75,200.00

   Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
13                       $122,000.00                       $154,000.00
1                        $127,000.00                       $161,000.00
30                       $132,000.00                       $173,000.00
38                       $120,000.00                       $163,000.00
19                       $130,000.00                       $168,000.00
12                       $135,000.00                       $162,000.00
8                        $143,000.00                       $194,000.00
43                       $108,000.00                       $124,000.00
```

[47]:
```python
# Plot the boxplot
fig, axes = plt.subplots(nrows=2, ncols=1,figsize=(8,10))
axes = axes.flatten()


box1=axes[0].
 ↪boxplot([ivy,engineering,lib_arts,party,state],patch_artist=True,vert=True)
```

```
# Customize the outline and fill color for boxplot
# https://stackoverflow.com/questions/41997493/python-matplotlib-boxplot-color
for box in box1['boxes']:
    # change outline color
    box.set(color='white', linewidth=2)
    # change fill color
    box.set(facecolor = '#d17f79',alpha=0.5 )

axes[0].text(0.5,80, 'School Type vs. Starting Median Salaries',␣
↪color="#696763",fontweight="bold",size=15)
axes[0].yaxis.set_ticks_position('none')
axes[0].xaxis.set_ticks_position('none')

# Set a general title for the whole plot
axes[0].text(0.5,85, 'Engineering Graduates Make Significantly More Than Most␣
↪Other Graduates  ', fontweight='bold',size=20)

# Add annotation on the first plot
axes[0].annotate('Match the second graph below', xy=(2.3, 60), xytext=(2.34,␣
↪70),
            arrowprops=dict(color='#f0b14d',arrowstyle='->'),␣
↪fontsize=11,color="#f0b14d")
# Unit for y axis
axes[0].text(-0.8,74.5,"salaries in 1000s")
axes[0].spines['right'].set_visible(False)
axes[0].spines['top'].set_visible(False)
# Change the boxplot color
colors = ['#d17f79','#f0b14d','#d17f79','#d17f79','#d17f79']
for item in ['boxes', 'fliers', 'medians', 'means']:
    for sub_item,color in zip(box1[item], colors):
        plt.setp(sub_item, color=color)

# Change the xticks names
plt.sca(axes[0])
plt.xticks([1,2,3,4,5], ['Ivy', 'Engineering', 'Lib_Arts',"Party","State"])



# Second plot (barplot)
axes[1].barh(top_ten_pay_df["Undergraduate Major"], top_ten_pay_df["Starting␣
↪Median Salary"],height=0.5,edgecolor="#a37731",color="#f0b14d",alpha=0.5)
axes[1].text(25,-2,"Starting median salary (1000s)")
axes[1].text(0,8, 'Top Eight Degrees with Highest Starting Salaries',␣
↪color="#696763",fontweight="bold",size=15)
axes[1].yaxis.set_ticks_position('none')
```

```
axes[1].xaxis.set_ticks_position('none')
#axes[1].set_yticklabels(top_ten_pay_df["Undergraduate Major"],ha='left')
# Add the annotations
axes[1].plot([80,80],[-0.45,6.3],  c='#7d5019')
axes[1].text(92,5,"7 out of 8 are", horizontalalignment='center',size=16,color="#7d5019")
axes[1].text(97,4.5,"engineering related", horizontalalignment='center',size=16,color="#7d5019")
axes[1].spines['right'].set_visible(False)
axes[1].spines['top'].set_visible(False)
# for spine in plt.gca().spines.values():
#     spine.set_visible(False)

#     for i in range(4):
#     axes[i].spines['left'].set_visible(False)
#     axes[i].spines['top'].set_visible(False)
#     axes[i].spines['right'].set_visible(False)
#     axes[i].spines['bottom'].set_linewidth(.5)
#     axes[i].yaxis.set_ticks_position('none')
#     axes[i].xaxis.set_ticks_position('none')



plt.savefig("Salary.png")
```



```
[ ]:
```