P3_en

December 24, 2018

```
In [1]: from urllib import request
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        sns.set()
```

```
In [2]: request.urlretrieve ("https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/cou
        hubble = pd.read_csv("hubble.csv")
```

```
In [3]: hubble.describe()
```

```
Out[3]:           distance   recession_velocity
        count   24.000000            24.000000
        mean     0.911375           373.125000
        std      0.645496           371.254666
        min      0.032000          -220.000000
        25%      0.406250           165.000000
        50%      0.900000           295.000000
        75%      1.175000           537.500000
        max      2.000000          1090.000000
```

```
In [4]: from sklearn.linear_model import LinearRegression

        X = hubble.distance.values.reshape(-1,1)
        Y = hubble.recession_velocity

        lr = LinearRegression()
        lr.fit(X, Y)
```

```
Out[4]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
                 normalize=False)
```

```
In [5]: BIGGER_SIZE = 15

        plt.rc('font', size=BIGGER_SIZE)           # controls default text sizes
        plt.rc('axes', titlesize=BIGGER_SIZE)      # fontsize of the axes title
        plt.rc('axes', labelsize=BIGGER_SIZE)      # fontsize of the x and y labels
        plt.rc('xtick', labelsize=BIGGER_SIZE)     # fontsize of the tick labels
```

```
plt.rc('ytick', labelsize=BIGGER_SIZE)      # fontsize of the tick labels
plt.rc('legend', fontsize=BIGGER_SIZE)       # legend fontsize
plt.rc('figure', titlesize=BIGGER_SIZE)    # fontsize of the figure title

plt.figure()
plt.title("Data from Hubble")
plt.scatter(X, Y,  color='black', label="data")
plt.plot(X, lr.predict(X), color='red', linewidth=3, label="fitted line")
plt.xlabel("Distance")
plt.ylabel("Recession velocity")
plt.legend()
plt.show()
```