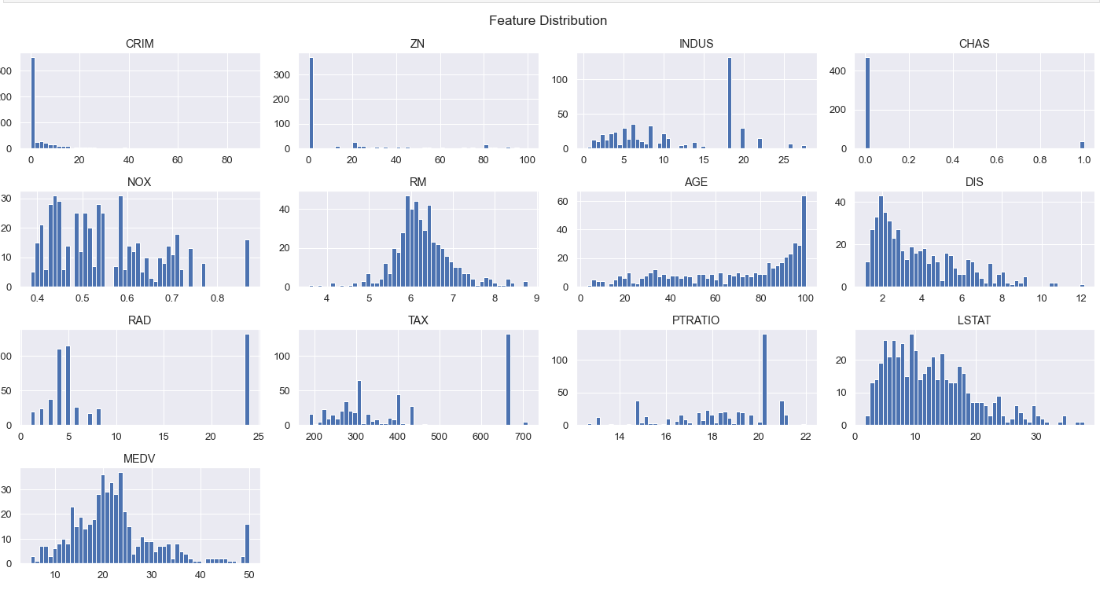
**Data Visualization**

df**.**hist(bins**=**50, figsize**=**(20,10))

plt**.**suptitle('Feature Distribution', x**=**0.5, y**=**1.02, ha**=**'center', fontsize**=**'large')

plt**.**tight\_layout()

plt**.**show()

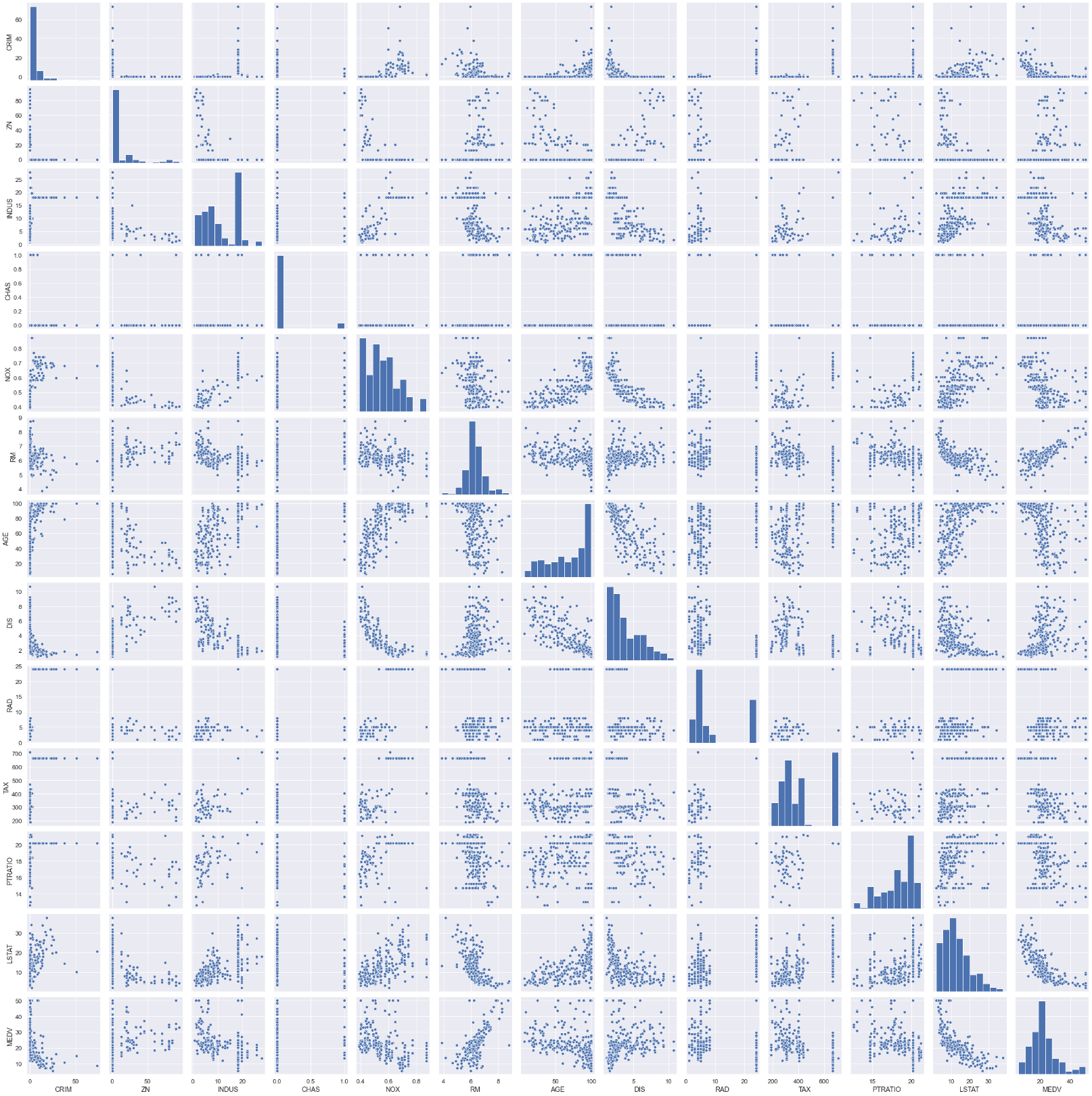


plt**.**figure(figsize**=**(20,20))

plt**.**suptitle('Pairplots of features', x**=**0.5, y**=**1.02, ha**=**'center', fontsize**=**'large')

sns**.**pairplot(df**.**sample(250))

plt**.**show()

****

**Task 4: Generate Descriptive Statistics and Visualizations**

In [9]:

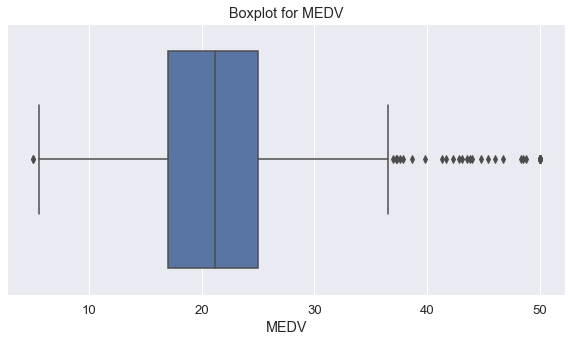
*#For the "Median value of owner-occupied homes" provide a boxplot*

plt**.**figure(figsize**=**(10,5))

sns**.**boxplot(x**=**df**.**MEDV)

plt**.**title("Boxplot for MEDV")

plt**.**show()

****

Note: Outliers after third quartile.

In [10]:

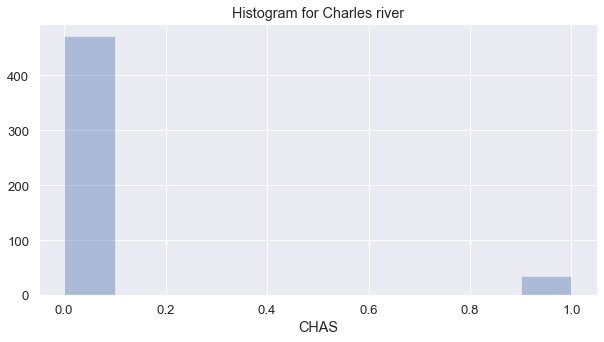
*#Provide a histogram for the Charles river variable*

plt**.**figure(figsize**=**(10,5))

sns**.**distplot(a**=**df**.**CHAS,bins**=**10, kde**=False**)

plt**.**title("Histogram for Charles river")

plt**.**show()

****

Note: Majority tracts don't bound Charles River

In [11]:

*#Provide a boxplot for the MEDV variable vs the AGE variable.*

*#(Discretize the age variable into three groups of 35 years and younger,*

*#between 35 and 70 years and 70 years and older)*

df**.**loc[(df["AGE"] **<=** 35),'age\_group'] **=** '35 years and younger'

df**.**loc[(df["AGE"] **>** 35) **&** (df["AGE"]**<**70),'age\_group'] **=** 'between 35 and 70 years'

df**.**loc[(df["AGE"] **>=** 70),'age\_group'] **=** '70 years and older'

In [12]:

df

Out[12]:

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** | **LSTAT** | **MEDV** | **age\_group** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.00632 | 18.0 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.0900 | 1 | 296 | 15.3 | 4.98 | 24.0 | between 35 and 70 years |
| **1** | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 9.14 | 21.6 | 70 years and older |
| **2** | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 4.03 | 34.7 | between 35 and 70 years |
| **3** | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 2.94 | 33.4 | between 35 and 70 years |
| **4** | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 5.33 | 36.2 | between 35 and 70 years |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **501** | 0.06263 | 0.0 | 11.93 | 0 | 0.573 | 6.593 | 69.1 | 2.4786 | 1 | 273 | 21.0 | 9.67 | 22.4 | between 35 and 70 years |
| **502** | 0.04527 | 0.0 | 11.93 | 0 | 0.573 | 6.120 | 76.7 | 2.2875 | 1 | 273 | 21.0 | 9.08 | 20.6 | 70 years and older |
| **503** | 0.06076 | 0.0 | 11.93 | 0 | 0.573 | 6.976 | 91.0 | 2.1675 | 1 | 273 | 21.0 | 5.64 | 23.9 | 70 years and older |
| **504** | 0.10959 | 0.0 | 11.93 | 0 | 0.573 | 6.794 | 89.3 | 2.3889 | 1 | 273 | 21.0 | 6.48 | 22.0 | 70 years and older |
| **505** | 0.04741 | 0.0 | 11.93 | 0 | 0.573 | 6.030 | 80.8 | 2.5050 | 1 | 273 | 21.0 | 7.88 | 11.9 | 70 years and older |

506 rows × 14 columns

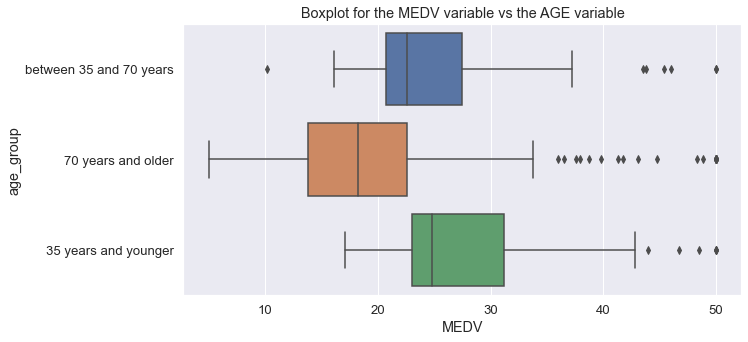
In [13]:

plt**.**figure(figsize**=**(10,5))

sns**.**boxplot(x**=**df**.**MEDV, y**=**df**.**age\_group, data**=**df)

plt**.**title("Boxplot for the MEDV variable vs the AGE variable")

plt**.**show()

****

Note: 35 years or younger group pays the highest median house price while above 70s are shifting to cheaper houses

In [14]:

*#Provide a scatter plot to show the relationship between Nitric oxide concentrations and*

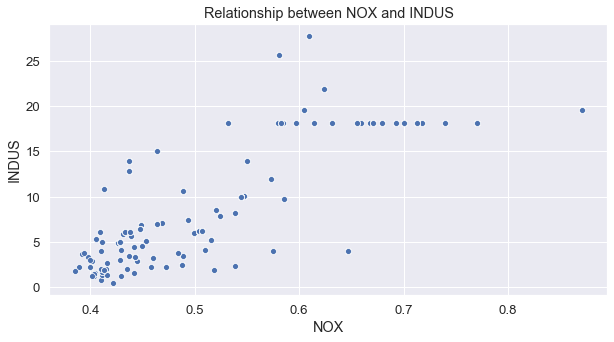
*#the proportion of non-retail business acres per town. What can you say about the relationship?*

plt**.**figure(figsize**=**(10,5))

sns**.**scatterplot(x**=**df**.**NOX, y**=**df**.**INDUS, data**=**df)

plt**.**title("Relationship between NOX and INDUS")

plt**.**show()

****

Note: There seems to be a linear relationship till NOX=0.6

In [15]:

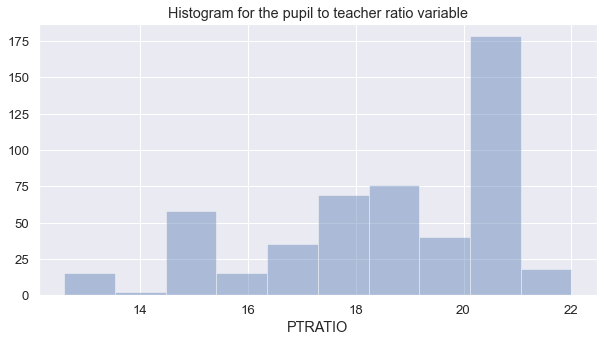
*#Create a histogram for the pupil to teacher ratio variable*

plt**.**figure(figsize**=**(10,5))

sns**.**distplot(a**=**df**.**PTRATIO,bins**=**10, kde**=False**)

plt**.**title("Histogram for the pupil to teacher ratio variable")

plt**.**show()

****

Note: Pupil to teacher ratio is highest at 20-21 range.

**Task 5: Use the appropriate tests to answer the questions provided**

In [16]:

df

Out[16]:

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** | **LSTAT** | **MEDV** | **age\_group** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.00632 | 18.0 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.0900 | 1 | 296 | 15.3 | 4.98 | 24.0 | between 35 and 70 years |
| **1** | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 9.14 | 21.6 | 70 years and older |
| **2** | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 4.03 | 34.7 | between 35 and 70 years |
| **3** | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 2.94 | 33.4 | between 35 and 70 years |
| **4** | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 5.33 | 36.2 | between 35 and 70 years |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **501** | 0.06263 | 0.0 | 11.93 | 0 | 0.573 | 6.593 | 69.1 | 2.4786 | 1 | 273 | 21.0 | 9.67 | 22.4 | between 35 and 70 years |
| **502** | 0.04527 | 0.0 | 11.93 | 0 | 0.573 | 6.120 | 76.7 | 2.2875 | 1 | 273 | 21.0 | 9.08 | 20.6 | 70 years and older |
| **503** | 0.06076 | 0.0 | 11.93 | 0 | 0.573 | 6.976 | 91.0 | 2.1675 | 1 | 273 | 21.0 | 5.64 | 23.9 | 70 years and older |
| **504** | 0.10959 | 0.0 | 11.93 | 0 | 0.573 | 6.794 | 89.3 | 2.3889 | 1 | 273 | 21.0 | 6.48 | 22.0 | 70 years and older |
| **505** | 0.04741 | 0.0 | 11.93 | 0 | 0.573 | 6.030 | 80.8 | 2.5050 | 1 | 273 | 21.0 | 7.88 | 11.9 | 70 years and older |

506 rows × 14 columns

#### Is there a significant difference in median value of houses bounded by the Charles river or not? (T-test for independent samples)

Null Hypothesis(𝐻0): Both average MEDV are the same

Alternative Hypothesis(𝐻1): Both average MEDV are NOT the same

#### Is there a significant difference in median value of houses bounded by the Charles river or not? (T-test for independent samples)

Null Hypothesis(𝐻0): Both average MEDV are the same

Alternative Hypothesis(𝐻1): Both average MEDV are NOT the same

df["CHAS"]**.**value\_counts()

Out[17]:

0 471

1 35

Name: CHAS, dtype: int64

In [18]:

a **=** df[df["CHAS"] **==** 0]["MEDV"]

a

Out[18]:

0 24.0

1 21.6

2 34.7

3 33.4

4 36.2

...

501 22.4

502 20.6

503 23.9

504 22.0

505 11.9

Name: MEDV, Length: 471, dtype: float64

In [19]:

b **=** df[df["CHAS"] **==** 1]["MEDV"]

b

Out[19]:

142 13.4

152 15.3

154 17.0

155 15.6

160 27.0

162 50.0

163 50.0

208 24.4

209 20.0

210 21.7

211 19.3

212 22.4

216 23.3

218 21.5

219 23.0

220 26.7

221 21.7

222 27.5

234 29.0

236 25.1

269 20.7

273 35.2

274 32.4

276 33.2

277 33.1

282 46.0

283 50.0

356 17.8

357 21.7

358 22.7

363 16.8

364 21.9

369 50.0

370 50.0

372 50.0

Name: MEDV, dtype: float64

In [20]:

scipy**.**stats**.**ttest\_ind(a,b,axis**=**0,equal\_var**=True**)

Out[20]:

Ttest\_indResult(statistic=-3.996437466090509, pvalue=7.390623170519905e-05)

Since p-value more than alpha value of 0.05, we failed to reject null hypothesis since there is NO statistical significance.

#### Is there a difference in Median values of houses (MEDV) for each proportion of owner occupied units built prior to 1940 (AGE)? (ANOVA)

In [21]:

df["AGE"]**.**value\_counts()

Out[21]:

100.0 43

96.0 4

98.2 4

95.4 4

97.9 4

..

47.6 1

92.7 1

13.9 1

58.4 1

40.1 1

Name: AGE, Length: 356, dtype: int64

In [22]:

df**.**loc[(df["AGE"] **<=** 35),'age\_group'] **=** '35 years and younger'

df**.**loc[(df["AGE"] **>** 35) **&** (df["AGE"]**<**70),'age\_group'] **=** 'between 35 and 70 years'

df**.**loc[(df["AGE"] **>=** 70),'age\_group'] **=** '70 years and older'

In [23]:

df

Out[23]:

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** | **LSTAT** | **MEDV** | **age\_group** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0.00632 | 18.0 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.0900 | 1 | 296 | 15.3 | 4.98 | 24.0 | between 35 and 70 years |
| **1** | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242 | 17.8 | 9.14 | 21.6 | 70 years and older |
| **2** | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242 | 17.8 | 4.03 | 34.7 | between 35 and 70 years |
| **3** | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222 | 18.7 | 2.94 | 33.4 | between 35 and 70 years |
| **4** | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222 | 18.7 | 5.33 | 36.2 | between 35 and 70 years |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **501** | 0.06263 | 0.0 | 11.93 | 0 | 0.573 | 6.593 | 69.1 | 2.4786 | 1 | 273 | 21.0 | 9.67 | 22.4 | between 35 and 70 years |
| **502** | 0.04527 | 0.0 | 11.93 | 0 | 0.573 | 6.120 | 76.7 | 2.2875 | 1 | 273 | 21.0 | 9.08 | 20.6 | 70 years and older |
| **503** | 0.06076 | 0.0 | 11.93 | 0 | 0.573 | 6.976 | 91.0 | 2.1675 | 1 | 273 | 21.0 | 5.64 | 23.9 | 70 years and older |
| **504** | 0.10959 | 0.0 | 11.93 | 0 | 0.573 | 6.794 | 89.3 | 2.3889 | 1 | 273 | 21.0 | 6.48 | 22.0 | 70 years and older |
| **505** | 0.04741 | 0.0 | 11.93 | 0 | 0.573 | 6.030 | 80.8 | 2.5050 | 1 | 273 | 21.0 | 7.88 | 11.9 | 70 years and older |

506 rows × 14 columns

State the hypothesis

* µµµ𝐻0:µ\_1=µ\_2=µ\_3 (the three population means are equal)
* 𝐻1: At least one of the means differ

In [24]:

low **=** df[df["age\_group"] **==** '35 years and younger']["MEDV"]

mid **=** df[df["age\_group"] **==** 'between 35 and 70 years']["MEDV"]

high **=** df[df["age\_group"] **==** '70 years and older']["MEDV"]

In [25]:

f\_stats, p\_value **=** scipy**.**stats**.**f\_oneway(low,mid,high,axis**=**0)

In [26]:

print("F-Statistic={0}, P-value={1}"**.**format(f\_stats,p\_value))

F-Statistic=36.40764999196599, P-value=1.7105011022702984e-15

Since p-value more than alpha value of 0.05, we failed to reject null hypothesis since there is NO statistical significance.

#### Can we conclude that there is no relationship between Nitric oxide concentrations and proportion of non-retail business acres per town? (Pearson Correlation)

State the hypothesis

* 𝐻0: NOX is not correlated with INDUS
* 𝐻1: NOX is correlated with INDUS

In [27]:

pearson,p\_value **=** scipy**.**stats**.**pearsonr(df["NOX"],df["INDUS"])

In [28]:

print("Pearson Coefficient value={0}, P-value={1}"**.**format(pearson,p\_value))

Pearson Coefficient value=0.7636514469209154, P-value=7.913361061236894e-98

Since the p-value (Sig. (2-tailed) < 0.05, we reject the Null hypothesis and conclude that there exists a relationship between Nitric Oxide and non-retail business acres per town.

#### What is the impact of an additional weighted distance to the five Boston employment centres on the median value of owner occupied homes? (Regression analysis)

State Hypothesis

Null Hypothesis: weighted distances to five Boston employment centres are not related to median value

Alternative Hypothesis: weighted distances to five Boston employment centres are related to median value

In [29]:

df**.**columns

Out[29]:

Index(['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX', 'PTRATIO', 'LSTAT', 'MEDV', 'age\_group'], dtype='object')

In [30]:

y **=** df['MEDV']

x **=** df['DIS']

In [31]:

x **=** sm**.**add\_constant(x)

In [32]:

results **=** sm**.**OLS(y,x)**.**fit()

In [33]:

results**.**summary()

Out[33]:

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results  **Dep. Variable:** | MEDV | **R-squared:** | 0.062 |
| **Model:** | OLS | **Adj. R-squared:** | 0.061 |
| **Method:** | Least Squares | **F-statistic:** | 33.58 |
| **Date:** | Tue, 03 Nov 2020 | **Prob (F-statistic):** | 1.21e-08 |
| **Time:** | 10:00:54 | **Log-Likelihood:** | -1823.9 |
| **No. Observations:** | 506 | **AIC:** | 3652. |
| **Df Residuals:** | 504 | **BIC:** | 3660. |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | 18.3901 | 0.817 | 22.499 | 0.000 | 16.784 | 19.996 |
| **DIS** | 1.0916 | 0.188 | 5.795 | 0.000 | 0.722 | 1.462 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 139.779 | **Durbin-Watson:** | 0.570 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 305.104 |
| **Skew:** | 1.466 | **Prob(JB):** | 5.59e-67 |
| **Kurtosis:** | 5.424 | **Cond. No.** | 9.32 |

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

In [34]:

np**.**sqrt(0.062) *##Pearson Coeffiecent valuea*

Out[34]:

0.24899799195977465

The square root of R-squared is 0.25, which implies weak correlation between both features

### Correlation

In [35]:

df**.**corr()

Out[35]:

|  | **CRIM** | **ZN** | **INDUS** | **CHAS** | **NOX** | **RM** | **AGE** | **DIS** | **RAD** | **TAX** | **PTRATIO** | **LSTAT** | **MEDV** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CRIM** | 1.000000 | -0.200469 | 0.406583 | -0.055892 | 0.420972 | -0.219247 | 0.352734 | -0.379670 | 0.625505 | 0.582764 | 0.289946 | 0.455621 | -0.388305 |
| **ZN** | -0.200469 | 1.000000 | -0.533828 | -0.042697 | -0.516604 | 0.311991 | -0.569537 | 0.664408 | -0.311948 | -0.314563 | -0.391679 | -0.412995 | 0.360445 |
| **INDUS** | 0.406583 | -0.533828 | 1.000000 | 0.062938 | 0.763651 | -0.391676 | 0.644779 | -0.708027 | 0.595129 | 0.720760 | 0.383248 | 0.603800 | -0.483725 |
| **CHAS** | -0.055892 | -0.042697 | 0.062938 | 1.000000 | 0.091203 | 0.091251 | 0.086518 | -0.099176 | -0.007368 | -0.035587 | -0.121515 | -0.053929 | 0.175260 |
| **NOX** | 0.420972 | -0.516604 | 0.763651 | 0.091203 | 1.000000 | -0.302188 | 0.731470 | -0.769230 | 0.611441 | 0.668023 | 0.188933 | 0.590879 | -0.427321 |
| **RM** | -0.219247 | 0.311991 | -0.391676 | 0.091251 | -0.302188 | 1.000000 | -0.240265 | 0.205246 | -0.209847 | -0.292048 | -0.355501 | -0.613808 | 0.695360 |
| **AGE** | 0.352734 | -0.569537 | 0.644779 | 0.086518 | 0.731470 | -0.240265 | 1.000000 | -0.747881 | 0.456022 | 0.506456 | 0.261515 | 0.602339 | -0.376955 |
| **DIS** | -0.379670 | 0.664408 | -0.708027 | -0.099176 | -0.769230 | 0.205246 | -0.747881 | 1.000000 | -0.494588 | -0.534432 | -0.232471 | -0.496996 | 0.249929 |
| **RAD** | 0.625505 | -0.311948 | 0.595129 | -0.007368 | 0.611441 | -0.209847 | 0.456022 | -0.494588 | 1.000000 | 0.910228 | 0.464741 | 0.488676 | -0.381626 |
| **TAX** | 0.582764 | -0.314563 | 0.720760 | -0.035587 | 0.668023 | -0.292048 | 0.506456 | -0.534432 | 0.910228 | 1.000000 | 0.460853 | 0.543993 | -0.468536 |
| **PTRATIO** | 0.289946 | -0.391679 | 0.383248 | -0.121515 | 0.188933 | -0.355501 | 0.261515 | -0.232471 | 0.464741 | 0.460853 | 1.000000 | 0.374044 | -0.507787 |
| **LSTAT** | 0.455621 | -0.412995 | 0.603800 | -0.053929 | 0.590879 | -0.613808 | 0.602339 | -0.496996 | 0.488676 | 0.543993 | 0.374044 | 1.000000 | -0.737663 |
| **MEDV** | -0.388305 | 0.360445 | -0.483725 | 0.175260 | -0.427321 | 0.695360 | -0.376955 | 0.249929 | -0.381626 | -0.468536 | -0.507787 | -0.737663 | 1.000000 |

In [36]:

plt**.**figure(figsize**=**(16,9))

sns**.**heatmap(df**.**corr(),cmap**=**"coolwarm",annot**=True**,fmt**=**'.2f',linewidths**=**2, cbar**=False**)

plt**.**show()

