DSCI 101 Midterm Exam

ⓘ This is a preview of the published version of the quiz

Started: Jun 21 at 6:41pm

Quiz Instructions

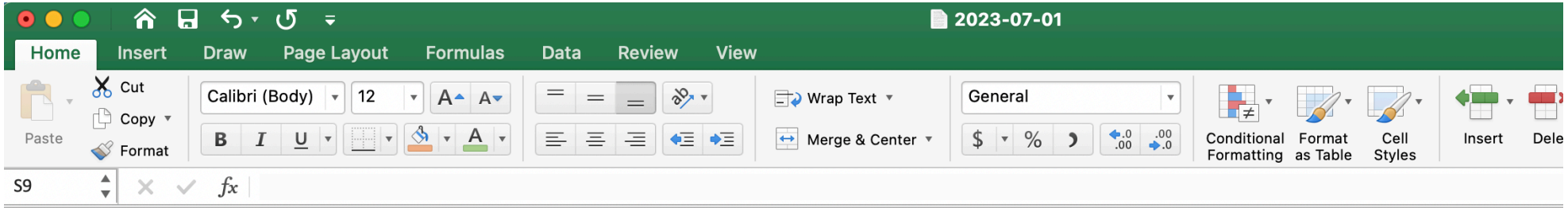
Exam rules:

- You may access any course material, including slides, demos, notes, and lab assignments.
- You may access other related books or reference material you can find.
- You may also passively use the internet. You may search the internet for material, but you cannot post a question and ask for help, including ChatGPT or other AI tools for help.
- **Your timer will start once you click Take the Quiz.**

You are working on a project to understand the compensation level of government employees in Texas. You have the following overarching questions:

- What is the median salary of full-time government employees in recent years?
- How does the median salary differ based on various factors, such as demographics, job types, employment experiences, etc?

To help answer these questions, you discovered this data set of compensation for Texas state employees published online (government employees' salary is public information). For your convenience, here is a snapshot of the data file in a CSV file:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	AGY	NAME	LASTNAME	FIRSTNAME	MI	JOBCLASS	JC TITLE	RACE	SEX	EMPTYPE	HIREDT	RATE	HRSWKD	MONTHLY	ANNUAL	STATENUM	
2	101	SENATE	ALLENSWOR	ANNE	R	7104	LEGISLATIVE	WHITE	FEMALE	URP - UNCLA	4/18/16	0	10	7083	84996	339273	
3	104	LEGISLATIVE	ALLENSWOR	ANNE	R	P070	ANALYST	WHITE	FEMALE	URP - UNCLA	1/1/22	0	30	5000	60000	339273	
4	101	SENATE	BELTRAN	AHITZA	G	7103	LEG. SERVICI	HISPANIC	FEMALE	URF - UNCLA	1/2/19	0	41	3750.72	45008.64	1164354	
5	101	SENATE	GILLIAM	STACEY	L	7101	LEG. OFFICIA	WHITE	FEMALE	URP - UNCLA	7/1/13	0	10	9188	110256	339371	
6	104	LEGISLATIVE	GILLIAM	STACEY	L	C160	COMMITTEE	WHITE	FEMALE	URP - UNCLA	11/1/17	0	30	9000	108000	339371	
7	101	SENATE	HERNANDEZ	PETE		7103	LEG. SERVICI	HISPANIC	MALE	URF - UNCLA	10/19/22	0	41	3476.66	41719.92	1562288	
8	101	SENATE	JONES	ALYSSA	N	7104	LEGISLATIVE	WHITE	FEMALE	URP - UNCLA	8/1/20	0	10	7345	88140	669977	
9	104	LEGISLATIVE	JONES	ALYSSA	N	P070	ANALYST	WHITE	FEMALE	URP - UNCLA	8/1/21	0	30	6000	72000	669977	
10	101	SENATE	KENNY	PAT		7104	LEGISLATIVE	WHITE	FEMALE	URF - UNCLA	2/3/05	0	41	5433.33	65199.96	37375	
11	101	SENATE	ROCHA	JONATHON	S	7103	LEG. SERVICI	HISPANIC	MALE	URF - UNCLA	9/12/22	0	41	3886.66	46639.92	1552496	
12	101	SENATE	ROCHA	MARIE	S	7103	LEG. SERVICI	HISPANIC	FEMALE	URF - UNCLA	5/1/03	0	41	4462.12	53545.44	152257	
13	101	SENATE	WHITE	MARK		7101	LEG. OFFICIA	WHITE	MALE	URF - UNCLA	1/23/04	0	40	6380	76560	152770	
14	101	SENATE	WHITE	MARK		7103	LEG. SERVICI	WHITE	MALE	URP - UNCLA	1/23/04	25	1	108.33	1299.96	152770	
15	241	COMPTROLL	SPECIA JR	JOHN	J	JD25	JUDGE, RETI	WHITE	MALE	URP - UNCLA	2/1/20	75.9615	29	9545.82	114549.84	59115	
16	212	OFFICE OF C	SPECIA JR	JOHN	J	3524	GENERAL CC	WHITE	MALE	CTP - CLASSI	9/1/18	81.04453	4	1404.77	16857.24	59115	
17	448	OFFICE OF IN	BURKE	LAUREN	L	3662	OMBUDSMA	BLACK	FEMALE	CRF - CLASSI	4/1/23	0	40	4500	54000	1155823	
18	529	HEALTH AND	BRIGHT	VONTI		1864	MANAGEME	BLACK	FEMALE	CRF - CLASSI	6/5/23	0	40	5416.67	65000.04	1553456	
19	696	TEXAS DEPAI	CALDWELL	SHAQUALA	N	4540	PAROLE OFF	BLACK	FEMALE	CRF - CLASSI	6/21/23	0	40	3475.35	41704.2	1250979	
20	529	HEALTH AND	CARTER	MONIQUE	A	1552	STAFF SERVI	OTHER	FEMALE	CRF - CLASSI	9/1/22	0	40	4738	56856	313432	
21	529	HEALTH AND	EVANS	KIMBERLY	R	1574	PROGRAM S	BLACK	FEMALE	CRF - CLASSI	9/1/22	0	40	4528.63	54343.56	164387	
22	696	TEXAS DEPAI	FORD	LISA	M	156	ADMINISTRA	WHITE	FEMALE	CRF - CLASSI	10/26/22	0	40	3088.48	37061.76	320456	

Each row in this data is one Texas state employee compensation record in the year 2023, and columns include variables whose names are mostly self-explanatory. Here are some important variables:

- NAME: the organization name which the employee belongs to
- LASTNAME, FIRSTNAME, and MI: name of the employee
- JC TITLE: job title
- RACE and SEX: basic demographic info of the employee
- EMPTYPE: employment type as classified/unclassified/regular/temporary etc.
- HIREDT: month/day/year of hire
- RATE: hourly rate if available
- HRSWKD: number of hours worked per week

- MONTHLY: monthly compensation
- ANNUAL: annual compensation

Question 1 4 pts

Finding this data set is corresponding to which stage of the data science pipeline?

Data wrangling and exploration.

Formulate questions and identify or collect data.

Modeling and model validation.

Interpret and communicate results.

Question 2 4 pts

Suppose this data set includes ALL Texas state employees currently employed, and you start your question:

- What is the median salary of full-time Texas state employees in 2023?

What type of data science questions is this?

Descriptive and exploratory.

None of them.

Predictive and decision-making.

○
Inferential and causal.

Question 3 4 pts

Suppose this data set only includes a SAMPLE of Texas state employees currently employed, and you start your question:

- What is an estimated range of the median salary of full-time Texas state employees in 2023?

What type of data science questions is this?

○
Descriptive and exploratory.

○
Predictive and decision-making.

○
None of them.

○
Inferential and causal.

Question 4 4 pts

You started data wrangling and exploration to help answer some of your questions. Which of the following lines of code will read this ***2023-07-01.csv*** file into Python Pandas and save it as a dataframe named ***tx_salary***?

○
tx_salary = pd.read_csv("2023-07-01.csv")

○
pd.read_csv("2023-07-01.csv")

○
tx_salary = pd.read_csv(2023-07-01.csv)

○
tx_salary == pd.read_csv("2023-07-01.csv")

Question 5 4 pts

Now that you have read in the data and saved this dataframe named ***tx_salary***, what does the following two lines of Python code returns?

1. type(tx_salary)
2. type("tx_salary")

○
1 returns dataframe, and 2 returns error message.

○
1 returns dataframe, and 2 returns string.

○
Both return dataframe.

○
Both return error message.

Question 6 4 pts

Which of the following df.method allows you to see all the column names and their corresponding data types in Python, an output looks like this:

AGY	int64
NAME	object
LASTNAME	object
FIRSTNAME	object
MI	object
JOBCLASS	object
JC TITLE	object
RACE	object
SEX	object
EMPTYPE	object
HIREDT	object
RATE	float64
HRSWKD	float64
MONTHLY	float64
ANNUAL	float64

○ tx_salary.astype

○ tx_salary.dtypes

○ tx_salary.shape

○ tx_salary.columns

Question 7 4 pts

Based on the previous question, what is the data type for the column ***HIREDT***?

○ Float numbers

○ Pandas datetime.

○ Integer numbers.

○ String or mixed type.

Question 8 4 pts

Suppose the data only includes those who are currently employed. Which is the correct code to find the record of TX governor in the data, shown as the row below?

1. tx_salary["JC TITLE"]=="GOVERNOR"

2. tx_salary[tx_salary["JC TITLE"]=="GOVERNOR"]

AGY	NAME	LASTNAME	FIRSTNAME	MI	JOBCLASS	JC TITLE	RACE	SEX	EMPTYPE	HIREDT	RATE	HRSWKD	MONTHLY	ANNUAL
301	OFFICE OF THE GOVERNOR	ABBOTT	GREGORY	W	G030	GOVERNOR	WHITE	MALE	ERF - EXEMPT REGULAR FULL-TIME	01/20/15	0.0	40.0	12812.5	153750.0

○
Both 1 and 2.

○
Neither 1 or 2.

○
2 only.

○
1 only.

Question 9 4 pts

How to sort the data by annual salary so that the highest ones are on top? See the top 3 highest annual salaries in the data below:

AGY	NAME	LASTNAME	FIRSTNAME	MI	JOBCLASS	JC TITLE	RACE	SEX	EMPTYPE	HIREDT	RATE	HRSWKD	MONTHLY	ANNUAL
323	TEACHER RETIREMENT SYSTEM	AUBY	JASE	R	C204	CHIEF INVESTMENT OFFICER	WHITE	MALE	ERF - EXEMPT REGULAR FULL-TIME	11/09/09	0.0	40.0	54166.66	649999.92
542	CANCER PREVENTION AND RESEARCH INSTITUTE OF TEXAS	LE BEAU	MICHELLE	NaN	C542	CHIEF SCIENTIFIC OFFICER	WHITE	FEMALE	ERF - EXEMPT REGULAR FULL-TIME	10/11/21	0.0	40.0	50737.50	608850.00
323	TEACHER RETIREMENT SYSTEM	GUTHRIE	BRIAN	K	E176	EXECUTIVE DIRECTOR	WHITE	MALE	ERF - EXEMPT REGULAR FULL-TIME	10/01/08	0.0	40.0	41666.66	499999.92

○
tx_salary.sort_values('ANNUAL', ascending=False)

○
tx_salary.sort_values('ANNUAL', ascending=True)

○
tx_salary.sort_index('ANNUAL', ascending=False)

○
tx_salary.sort_index(ascending=False)

Question 10 4 pts

Suppose a string method str.contains('substring') will return True if the substring can be matched in the input string, otherwise returns False. For example:

my_str = ['DSCI101midterm exam', 'is', 'so much', 'fun']

my_str.str.contains('exam') returns [True, False, False, False]

What does the following line of code return?

tx_salary[tx_salary['JC TITLE'].str.contains('ACTUARY')]['ANNUAL'].mean()

○
One row with all columns.

○
One column with some rows.

○
One column with all rows.

○
One number.

Question 11 4 pts

What does the following cell of code return?

```
tx_salary = tx_salary[(tx_salary['HRSWKD']>=40)&(tx_salary['ANNUAL']>=30000)]
tx_salary[tx_salary['HRSWKD']<40]
```

○ Empty dataframe with only column names but no rows.

○ Rows where hours worked per week is more than 40 hours.

○ Annual salary number for rows where hours worked per week is less than 40 hours.

○ Rows where hours worked per week is less than 40 hours.

Question 12 4 pts

You try the following code, and see the output below:

```
tx_salary_ft = tx_salary[tx_salary['HRSWKD']>=40]
tx_salary_ft_20 = tx_salary_ft[tx_salary_ft['HIREDT']>='1/1/20']
tx_salary_ft_20
```

AGY	NAME	LASTNAME	FIRSTNAME	MI	JOBCLASS	JC TITLE	RACE	SEX	EMPTYPE	HIREDT	RATE
101	SENATE	HERNANDEZ	PETE	NaN	7103	LEG. SERVICE/MAINTENANCE	HISPANIC	MALE	URF - UNCLASSIFIED REGULAR FULL-TIME	10/19/22	0.0
696	TEXAS DEPARTMENT OF CRIMINAL JUSTICE	FORD	LISA	M	0156	ADMINISTRATIVE ASSISTANT IV	WHITE	FEMALE	CRF - CLASSIFIED REGULAR FULL-TIME	10/26/22	0.0
529	HEALTH AND HUMAN SERVICES COMMISSION	LIGGINS	DEMETRIA	S	1323	INSPECTOR III	BLACK	FEMALE	CRF - CLASSIFIED REGULAR FULL-TIME	12/07/22	0.0
101	SENATE	ALBERS	FRANCES	N	7104	LEGISLATIVE PROFESSIONAL	WHITE	FEMALE	URF - UNCLASSIFIED REGULAR FULL-TIME	10/01/14	0.0
101	SENATE	ALMAGUER	FRANK	NaN	7106	LEGISLATIVE PROTECTIVE SERVICE	HISPANIC	MALE	URF - UNCLASSIFIED REGULAR FULL-TIME	11/10/14	0.0

Notice the 4th and 5th rows showing here with a hire date before 1/1/20, why and what is wrong here?

- ○ The two conditions for filtering should be combined with | (logical operator OR).
- ○ The two conditions for filtering should be combined with & (logical operator AND).
- ○ 'HIREDT' column is a string type in this case, so >= '1/1/20' does not compare dates correctly.

○
Need to filter on 'HIREDT' first, then filter on 'HRSWKD' second.

Question 13 4 pts

How to correctly filter on ***HIREDT*** column?

1. Use str.split to split the string by '/', extract the year and convert to integers.

2. Use pd.to_datetime to convert the 'HIREDT' column to datetime type, then can extract year as integers.

○
1 only.

○
Neither 1 or 2 will work.

○
2 only.

○
Either 1 or 2 could work.

Question 14 4 pts

Now suppose you have correctly filtered out all the full-time employees hired in 2020 or later and named this new dataframe as ***tx_salary_ft_20***. Next you want to do a sanity check to understand the relation between columns ***MONTHLY*** and ***ANNUAL***. What does the following code return?

```
np.sum(tx_salary_ft_20['MONTHLY']*12 == tx_salary_ft_20['ANNUAL'])
```

○
The number of rows whose annual salary is monthly salary times 12.

○
A dataframe with rows whose annual salary is monthly salary times 12.

○
The number of rows whose annual salary is NOT monthly salary times 12.

○
A dataframe with rows whose annual salary is NOT monthly salary times 12.

Question 15 4 pts

After investigating and reaching out to the data source, you come to the conclusion that the ***ANNUAL*** column is not to be trusted. Rather you will calculate the correct annual number by multiplying monthly salaries by 12, and save it as a column in ***tx_salary_ft_20*** as ***ANNUAL_CORRECT***. Which is the correct code to do this?

○
tx_salary_ft_20['ANNUAL_CORRECT'] = tx_salary_ft_20['MONTHLY']*12

○
tx_salary_ft_20['ANNUAL'] = tx_salary_ft_20['MONTHLY']*12

○
tx_salary_ft_20['ANNUAL_CORRECT'] = tx_salary_ft_20['MONTHLY']

○
none of them is correct, needs to use a for loop to calculate each monthly number times 12.

Question 16 4 pts

You want to check if there are any missing values in ***tx_salary_ft_20***, if so, you want to find how many missing values are there in each column. Which of the following code will return this?

○
tx_salary_ft_20.isna()

○
tx_salary_ft_20[tx_salary_ft_20.isna()]

○
tx_salary_ft_20.isna().any()

○
tx_salary_ft_20.isna().sum()

Question 17 4 pts

It looks like there are no missing values except the column ***MI*** for the middle initial, which you don't really care about. Next, you want to create one data visualization to explore the distribution of the monthly salary. Which of the following plots is the best choice?

○
A histogram.

○
A bar plot.

○
A scatter plot.

○
A line plot.

Question 18 4 pts

Which is the correct code to generate this output where the ***MONTHLY*** column shows the median monthly salary for each race?

RACE	MONTHLY
AM INDIAN	3816.650
ASIAN	5100.000
BLACK	3748.000
HISPANIC	3750.000
OTHER	3649.830
WHITE	3946.845

- tx_salary_ft_20[['MONTHLY', 'RACE']].groupby('RACE').median()
- tx_salary_ft_20[['MONTHLY', 'RACE']].groupby('RACE').mean()
- tx_salary_ft_20[['MONTHLY', 'RACE']].groupby('MONTHLY').mean()
- tx_salary_ft_20[['MONTHLY', 'RACE']].groupby('MONTHLY').median()

Question 19 4 pts

Which is the correct code to generate this exact output to show the median monthly salary for each race and gender?

SEX	FEMALE	MALE
RACE		
AM INDIAN	3816.65	3881.13
ASIAN	5205.74	4748.95
BLACK	3670.64	3776.91
HISPANIC	3602.00	3776.91
OTHER	3574.18	3776.91
WHITE	3843.18	4002.89

This is real data! Good job, Asian females:)

○ tx_salary_ft_20.pivot_table(index='RACE', columns='SEX', values='MONTHLY', aggfunc='median')

○ tx_salary_ft_20.pivot_table(index='SEX', columns='RACE', values='MONTHLY', aggfunc='median')

○ tx_salary_ft_20.pivot_table(index='RACE', columns='SEX', values='median', aggfunc='MONTHLY')

○
tx_salary_ft_20.pivot_table(index='RACE', columns='SEX', values='MONTHLY', aggfunc='mean')

Question 20 4 pts

The dataframe ***tx_salary_ft*** includes all the full-time employees that work for at least 40 hours per week, with the hire date back to several decades. To explore the median income by year, you want to add a new column in ***tx_salary_ft*** named ***HIREYR*** to code the year as integers correctly. Which is the correct code do this?

○
tx_salary_ft['HIREYR'] = tx_salary['HIREDT'].str.split('/')

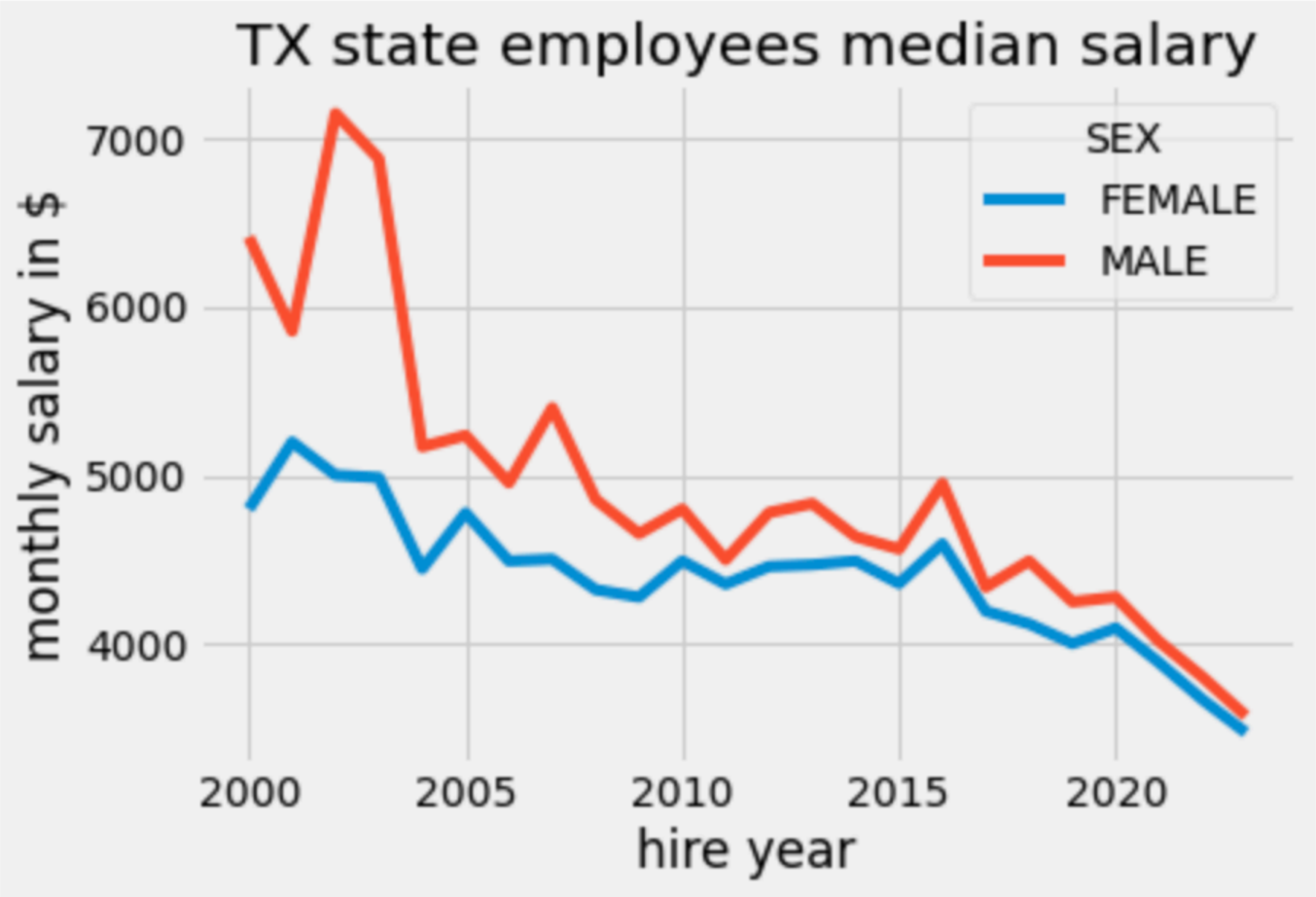
○
tx_salary_ft['HIREYR'] = pd.to_datetime(tx_salary['HIREDT'])

○
tx_salary_ft['HIREYR'] = pd.to_datetime(tx_salary['HIREDT']).dt.year

○
pd.to_datetime(tx_salary['HIREDT']).dt.year

Question 21 4 pts

You decide to only look at the salary for employees hired in 2000 and later. So you create a new data frame by filtering on the hire year between 2000 to 2023, named ***tx_salary_ft_yr***. You want to look at this trend by year and compare male and female employees, a plot looks like this:



Which is the correct code to create this plot (you can ignore adding title and label details)?

○
tx_salary_ft_yr.pivot_table(index='HIREYR', columns='SEX', values='MONTHLY', aggfunc='median').plot();

○ tx_salary_ft_yr.pivot_table(index= 'SEX', columns='HIREYR', values='MONTHLY', aggfunc='median').plot();

○ tx_salary_ft_df[['SEX', 'HIREYR', 'MONTHLY']].groupby('HIREYR').median().plot();

○ tx_salary_ft_df[['SEX', 'HIREYR', 'MONTHLY']].groupby('SEX').median().plot();

Question 22 4 pts

How would you find the top 10 TX government organizations that pay the highest average monthly salary? See the results below.

NAME	MONTHLY
TEXAS PERMANENT SCHOOL FUND CORPORATION	13003.836979
COMPTROLLER OF PUBLIC ACCOUNTS, JUDICIARY SECTION	12898.633505
TREASURY SAFEKEEPING TRUST COMPANY	10633.872073
CANCER PREVENTION AND RESEARCH INSTITUTE OF TEXAS	10307.882000
FIFTH COURT OF APPEALS DISTRICT	9765.014255
TENTH COURT OF APPEALS DISTRICT	9695.005000
TWELFTH COURT OF APPEALS DISTRICT	9604.923636
FOURTEENTH COURT OF APPEALS DISTRICT	9583.593784
SIXTH COURT OF APPEALS DISTRICT	9471.692727
FIRST COURT OF APPEALS DISTRICT	9367.909459

○ tx_salary_ft_df[['NAME', 'MONTHLY']].groupby('NAME').mean().head(10)

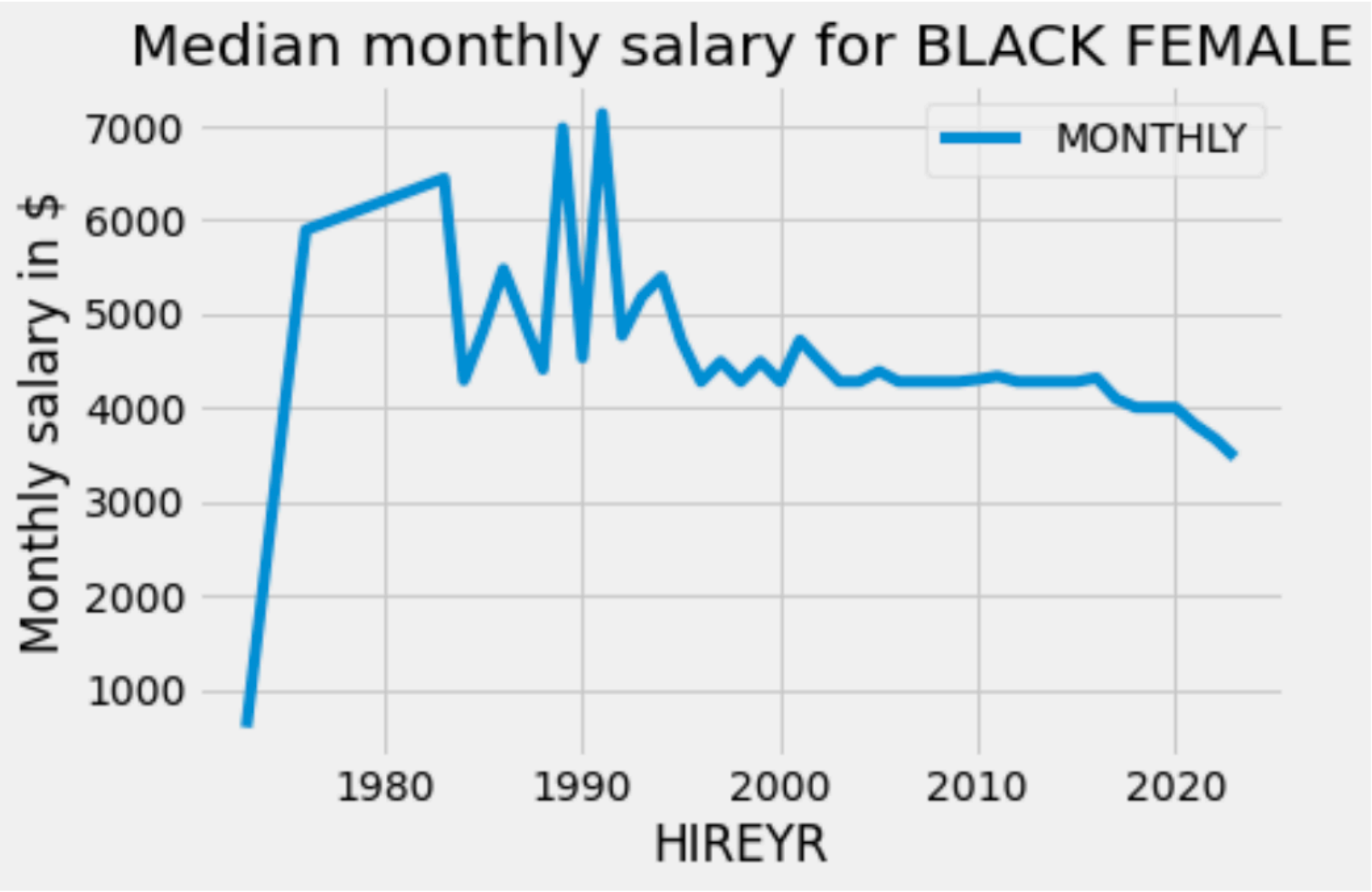
○ tx_salary_ft_df[['NAME', 'MONTHLY']].groupby('NAME').mean().sort_index().head(10)

○ tx_salary_ft_df[['NAME', 'MONTHLY']].groupby('NAME').mean().sort_values('MONTHLY', ascending=False).head(10)

○ tx_salary_ft_df[['NAME', 'MONTHLY']].groupby('NAME').count().head(10)

Question 23 4 pts

You would like to create a function for visualization. In particular, you would like to pick any demographic group (race and sex), and then look at the median salary for different hire years as a line plot. For example, if you pick black females, your function would create a plot like below:



In addition to the dataframe, what would be some important inputs for your function?

○
Hire year.

○
No inputs needed.

○
monthly salary.

○
Race and sex.

Question 24 4 pts

You want to investigate which job title has the most unequal pay for sex. Your plan is to find the median monthly salary for different job titles for both male and female groups, then use the ratio of female/male salary as a measure for unequal pay. See some of the results below:

SEX	FEMALE	MALE	FEMALE/MALE
JC TITLE			
CHIEF INVESTMENT OFFICER	26666.67	54166.660	0.49
PROG III	4582.55	7100.000	0.65
INVESTMENT ANLYST I	4301.16	6666.670	0.65
DATA ARCHITECT II	9000.00	13716.000	0.66
ASSISTANT DIRECTOR	7813.80	11687.505	0.67
EMERG MGT PROGRAM COORD V	5565.00	7979.180	0.70
CHAPLAIN III	4735.26	6579.410	0.72
DOCUMENT SERVICES TECH II	2338.70	3260.495	0.72
ASSISTANT GENERAL COUNSEL	7800.00	10875.005	0.72
RECORDS ANALYST III	3652.99	4988.630	0.73

Please select the correct code below in the correct order to create this dataframe.

1. my_df = tx_salary_ft_df[(tx_salary_ft_df['SEX'] == 'FEMALE') | (tx_salary_ft_df['SEX'] == 'MALE')]
2. my_df.sort_values(by = 'FEMALE/MALE')
3. my_df = tx_salary_ft_df['SEX', 'MONTHLY']. groupby('SEX').meadian()
4. my_df = tx_salary_ft_df.pivot_table(index='JC TITLE', columns='SEX', values='MONTHLY', aggfunc='median')
5. my_df = tx_salary_ft_df['JC TITLE', 'MONTHLY']. groupby("JC TITLE',').meadian()
6. my_df['FEMALE/MALE'] = my_df['FEMALE'] / my_df['MALE']

○
1 - 3 - 6 - 2

○
1 - 6 - 2

○
4 - 6 - 2

○
1 - 3 - 4 - 6 - 2

Question 25 4 pts

You further investigate the unequal pay issue by looking into the job title ***CHIEF INVESTMENT OFFICER***, which has the most unequal pay between male and female employees. You use code to get all the employees with that particular job title, and see the results below:

	AGY	NAME	LASTNAME	FIRSTNAME	MI	JOBCLASS	JC TITLE	RACE	SEX	EMPTYPE	HIREDT	RATE	HRSWKD	MONTHLY
17500	323	TEACHER RETIREMENT SYSTEM	AUBY	JASE	R	C204	CHIEF INVESTMENT OFFICER	WHITE	MALE	ERF - EXEMPT REGULAR FULL-TIME	11/09/09	0.0	40.0	54166.66
146219	930	TREASURY SAFEKEEPING TRUST COMPANY	ION	ANCA	M	1165	CHIEF INVESTMENT OFFICER	WHITE	FEMALE	URF - UNCLASSIFIED REGULAR FULL-TIME	08/24/05	0.0	40.0	26666.67

It turns out there are only two employees with that job title! Similarly, you looked up all the job titles listed in the Question above, and all of them have a small number of employees (mostly fewer than 20). You realized to better investigate gender unequal pay, you should focus on job titles that have a large number of employees. Which is the correct code to find a list of 50 job titles with the most number of employees?

○
top_title = tx_salary_ft['JC TITLE'].value_counts().sort_index().head(50)

○
top_title = tx_salary_ft['JC TITLE'].value_counts().head(50).index

○
top_title = tx_salary_ft['JC TITLE'].value_counts().sort_index()

○
top_title = tx_salary_ft['JC TITLE'].value_counts().index.head(20)

Not saved Submit Quiz