НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Факультет інформатики та обчислювальної
техніки

Пояснювальна записка

з дисципліни

“Ймовірнісні моделі та статистичне оцінювання в інформаційних системах”

Виконав:
Студент групи ЗПІ-зп01
Дишкант Л. Л.

Перевірив :
доцент кафедри ФІОТ
Ліхоузова Т. А.

Київ 2021

1. Загальна характеристика предметної області та постановка задачі

За концепцією В.Петті зарплата є грошовим виразом “ мінімуму засобів існування.”

У сучасній економічній теорії праця одночасно вважається фактором виробництва, а заробітна плата - ціною використання праці робітника. Прихильниками цієї концепції є відомі американські економісти П.Самуельсон, В.Нордгауз.

Закон України “ Про оплату праці” у преамбулі визначає відтворювальну та стимулюючу функції заробітної плати. Вчені доповнюють цей перелік соціальною, регелюючою та функцією формування платоспроможного попиту населення.

Відтворювальна - забезпечення працівників та членів їх сімей необхідними життєвими благами для відновлення робочої сили. Для того, щоб підтримувати своє життя,щоб виростити своїх дітей, які повинні його замінити на ринку праці. Крім того працівник може нести витрати для того, щоб розвинути свою робочу силу і отримати певну кваліфікацію.

Стимулююча - покликана заохотити працівника до постійного поліпшення якості та результатів власної праці через встановлення залежності її розміру від кількості та якості праці. Ця функція проявляється у структурній побудові заробітної плати на рівні додаткової заробітної плати та інших заохочувальних

виплат у формі надбавок, доплат та різноманітних премій.

Соціальну можна розглядати, як три напрямки:

- індикатор економічного розвитку держави;
- розподіл загального грошового фонду суспільства між державою, підприємцями та працівниками;
- прояв інтересів та можливість впливу різних верств суспільства на розподіл національного доходу.

Регулююча функція оплати праці характеризує оптимізацію розміщення робочої сили за регіонами, галузями господарства, підприємствами залежно від ринкової кон'юнктури.

Держава і підприємство встановлюють такі принципи диференціації заробітної плати працівників:

- величина заробітної плати залежить від складності праці, професійних навичок і кваліфікації робітника;
- від умов роботи, від її важкості, шкідливості для здоров'я (у важких і шкідливих оплачується вище);
- від результатів виробничої діяльності фірми в цілому.

Наразі в Україні більшість населення стрімко бідніє на фоні стрімкого підвищення тарифів за комунальні послуги (які часом перевищують зарплати багатьох українців і, тим більше, пенсії, особливо в

зимовий період) і високого рівня інфляції. Наявність в Україні значної кількості працюючих, які перебувають за межею бідності, загрожує соціальній стабільності, зменшує мотивацію працівників до продуктивної праці і породжує інші економічно-соціальні проблеми (тіньова економіка, злочинність, демографічні проблеми, заробітки за кордоном тощо). Тому дослідження динаміки рівня заробітної плати в Україні в контексті виконання нею своїх функцій є актуальним.

1.1 **Огляд предметної області**

З висновку щорічного дослідження провідної міжнародної консалтингової компанії “Mercer” випливає, що Україна у

2007 р продемонструвала найгірший розвиток рівня реальних зарплат в Європі. Тому одним з питань яке ми розглянемо:

Чи залежить заробітна плата в Україні від курсу долара?

В програмі діяльності Уряду “Український прорив: для людей, а не для політиків” зазначено, що гармонійний розвиток людини можливий за умови формування громадянського суспільства, забезпечення рівних прав та можливостей жінок і чоловіків, задоволення культурно-духовних потреб, високоякісної освіти та науки, сучасної медицини, безпечного довкілля, реалізації права на працю та гарантії соціального захисту. Другим питанням розглядатимемо:

Чи залежить заробітна плата від статі (чоловік чи жінка)?

1.2 **Огляд доступних джерел даних**

Маємо “Датасет” з даними середньомісячної заробітної плати щоквартально з 2015 по 2021 роки, а також з даними середньомісячної заробітної плати чоловіків та жінок окремо, дані про курс долара взяли з відкритих джерел інтернету (також поквартально з 2015 по 2021р середній) та додали до нашого “Датасету”.

Використаємо мову Python з додатковими бібліотеками та описову статистику для аналізу наших даних.

Так як ми будемо використовувати різні бібліотеки для розрахунків імпортуємо ті,що будуть нам потрібні. Читаем csv-файл для перевірки виведемо на екран перші п’ять стрічок (рис 1):

```
#читаємо csv-файл та виводимо перші 5 стрічок
dt = pd.read_csv('Zr.csv', sep=",", header = 1, index_col = False)
dt.head()
```

	code	attributes	period	average_all	salary_men	salary_woman	curs
0	1.0	У середньому по економіці	2015 Q1	3641	4238	3122	2339
1	2.0	Сільське господарство лісове господарство та р...	2015 Q1	2670	2762	2458	2339
2	2.1	сільське господарство	2015 Q1	2522	2592	2378	2339
3	3.0	Промисловість	2015 Q1	4236	4688	3434	2339
4	4.0	Будівництво	2015 Q1	2957	2993	2807	2339

рис. 1

Подивимося на статистичні оцінки:

- міри центральної тенденції (рис.2), де середнє - mean, кількість значень в відповідній колонці - count, медіана - 50%.

- міри варіативності (рис.2) :середньоквадратичне відхилення - std, розмах варіант -

min - max (мінімальне та максимальне значення відповідно), 75%-25% - міжквартильний розмах (більш стійкий до викидів).

```
# виведемо статистичні дані
dt[['average_all', 'salary_men', 'salary_woman', 'curs']].describe()
```

	average_all	salary_men	salary_woman	curs
count	675.000000	675.000000	675.000000	675.000000
mean	8743.315556	9678.589630	7845.134815	2598.480000
std	5251.913451	6121.940058	4301.252963	149.844432
min	1983.000000	2317.000000	1844.000000	2339.000000
25%	5280.000000	5834.500000	4834.000000	2526.000000
50%	7603.000000	8268.000000	7020.000000	2618.000000
75%	10713.500000	11793.500000	9672.500000	2731.000000
max	39439.000000	47465.000000	32116.000000	2827.000000

рис.2

Як бачимо з оцінок середнє значення заробітної плати у всіх перевищує медіану, що може попереджати про зсув розподілу вліво (лог.нормальне), а також може вказувати на викиди (великі значення, які не дуже типові більшості нашим даним). Перевіримо побудувавши гістограми (рис.3)б (рис.4) :

```
# побудуємо гістограми
dt[['average_all', 'salary_men', 'salary_woman', 'curs']].hist()
```

```
array([[<AxesSubplot:title={'center':'average_all'}>,
        <AxesSubplot:title={'center':'salary_men'}>],
       [<AxesSubplot:title={'center':'salary_woman'}>,
        <AxesSubplot:title={'center':'curs'}>]], dtype=object)
```

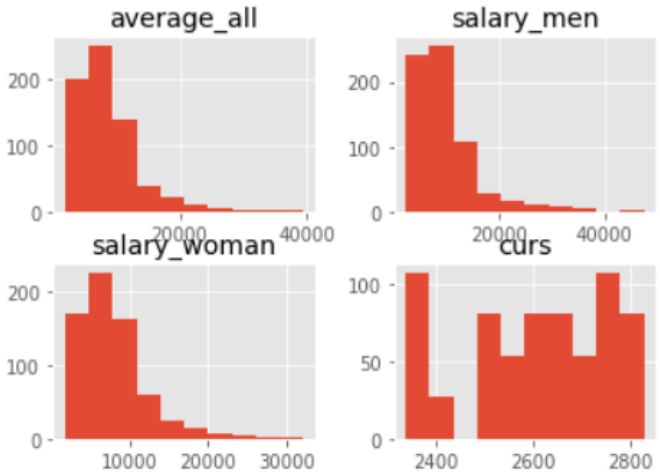


рис.3

```
# гістограми даних після логарифмування
np.log(dt[['average_all', 'salary_men', 'salary_woman', 'curs']]).hist()
```

```
array([[<AxesSubplot:title={'center':'average_all'}>,
        <AxesSubplot:title={'center':'salary_men'}>],
       [<AxesSubplot:title={'center':'salary_woman'}>,
        <AxesSubplot:title={'center':'curs'}>]], dtype=object)
```

рис.4

Побудуємо ядерні оцінки плотності наших даних до логарифмування :

```
my_density = gaussian_kde(dt['average_all'], bw_method = 0.1)

#график
x = linspace(min(dt['average_all']), max(dt['average_all']), 1000)
plot(x, my_density(x), 'g')
```

```
[<matplotlib.lines.Line2D at 0x1c5fd6666d0>]
```

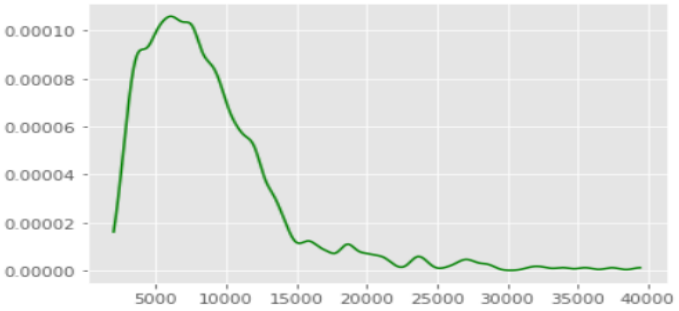


рис.5

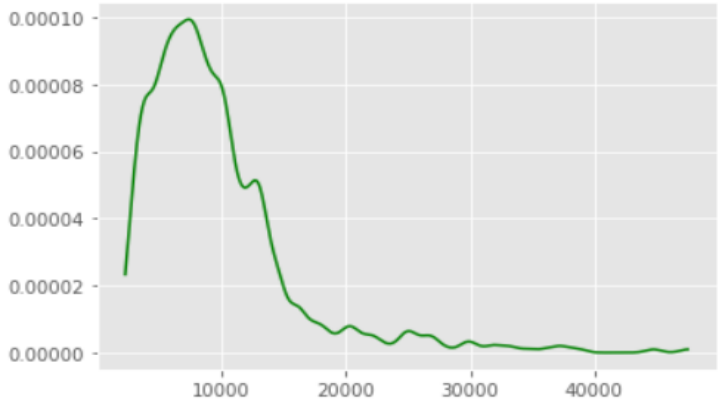


рис.6

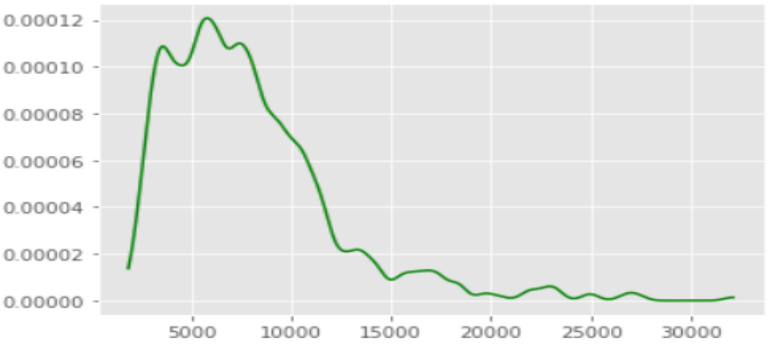


рис.7

Подивимося та проаналізуємо викиди, тому що аналізуємо дані з заробітною платою в галузях, побудуємо графік (рис.8). Знайдемо всі дані середньої заробітної плати в “авіаційному транспорті”, створимо

нову таблицю та побудуємо графік в залежності від часу, як змінювалися дані :

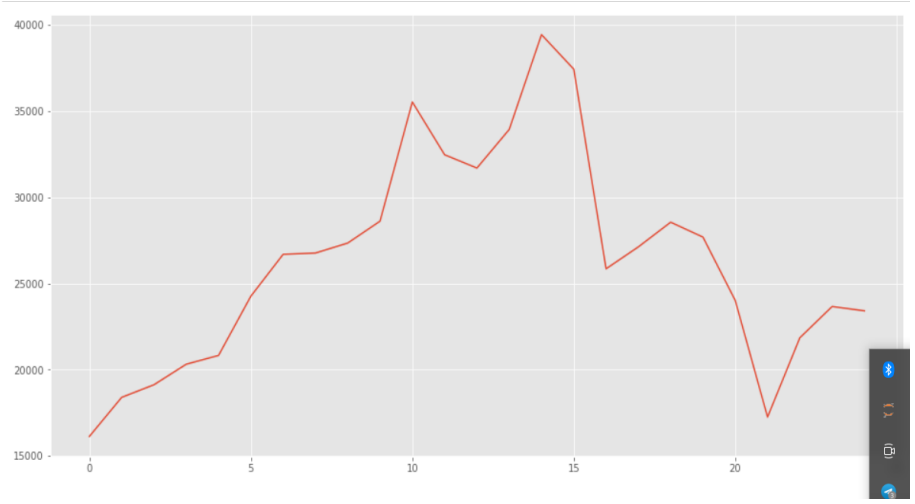


рис.8

Бачимо різкі спади (3 - 4) квартал 2019 р - (1-2)квартал 2020 року, наслідки, для галузі “авіаційний транспорт”, мір, які мали б стримати поширення інфекційної хвороби COVID19.

Проаналізуємо, ще дані галузі “Промисловість”, побудуємо графік (рис.9):

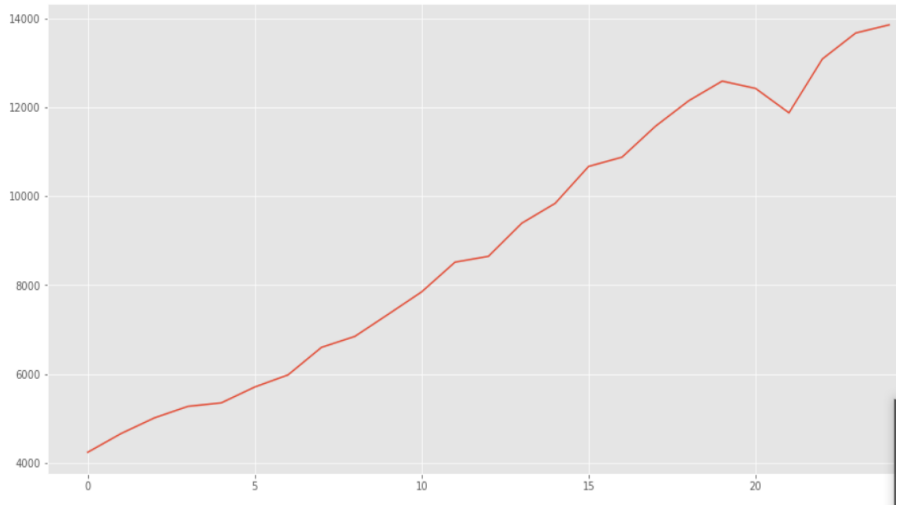


рис.9

Аналогічно попередньому графіку також маємо спад, менший, але і середні зарплати відповідно також менші.

1.3 **Постановка задачі**

Мета дослідження: проаналізувати дані, отримати відповіді на питання та зробити висновки.

1. Чи залежить заробітна плата від курсу долара?
2. Чи залежить заробітна плата від статі (чоловік чи жінка працює) ?
3. Чи має сезонність та чи зростає заробітна плата в Україні?

2. Вибір моделей

2.1 Визначення моделей, що можуть бути використані

Побудуємо лінійну та поліноміальну регресію.

Розглянемо на вибірці додатково Авторегресію та модель Аріма.

2.2 Вибір ознак, що будуть використані для аналізу

Перед побудовою моделей подивимося на теплову карту з матрицею кореляції (рис.10) та (рис.11):

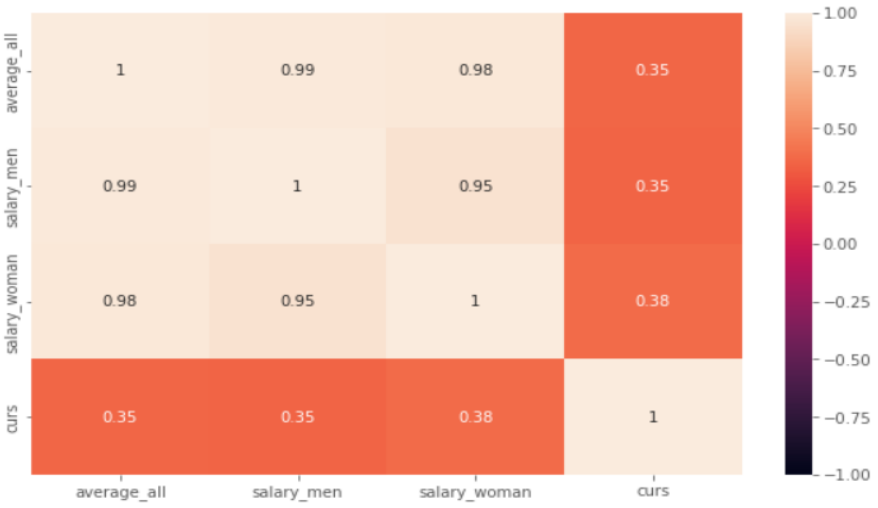


рис.10

рис.11

Побудуємо графік та діаграму розсіювання середньої заробітної плати в залежності від періоду ((рис.12),(рис.13)) без викидів (ми прибрали з таблиці з даними, дані тих галузей, зарплата в яких була не типова (наприклад - “авіаційний транспорт”) :

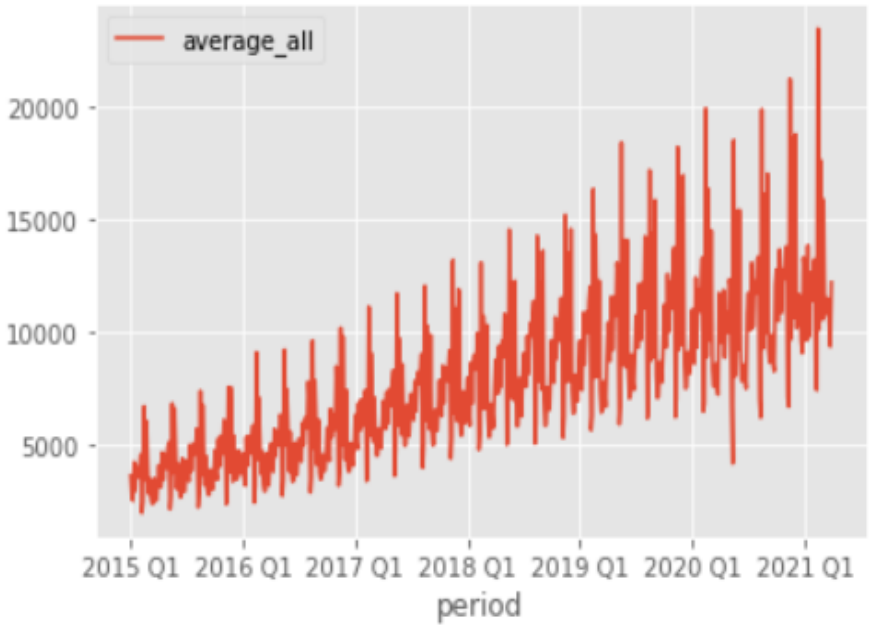


рис.12

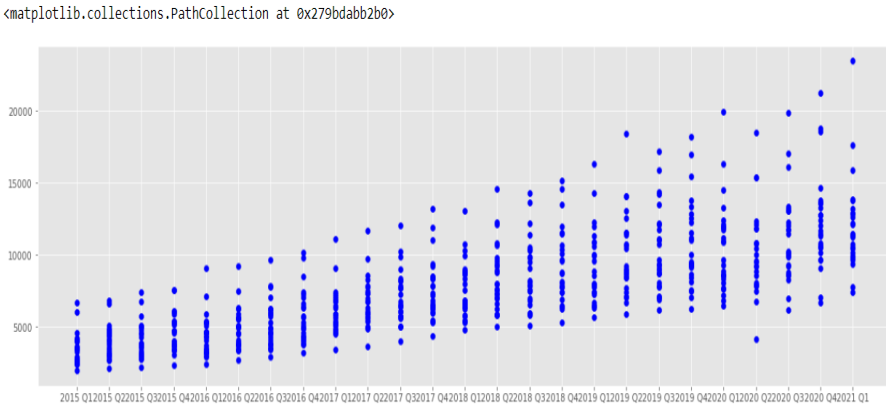


рис.13

Як бачимо з теплової карти, залежність між заробітною платою та курсом долара є , не дуже висока - 0.35, не зовсім схожа на лінійну з діаграм розсіювання. В залежності від періоду схожа на залежність близьку до лінійної (рис.12).

Але це для всіх галузей ми розглядали, зробимо вибірку з трьох галузей та побудуємо також теплову карту і діаграми розсіювання.

Зробили вибірку та створили нову таблицю з даними про :

середній курс долара, “сільське господарство” (код 02.1), “Освіта” (код 14), “Охорона здоров’я та надання соціальної допомоги”(код 15) за період з 2015 р - 2021 р (1 квартал) поквартально. Робимо описову статистику ((рис.14), (рис.15), (рис.16)), будуємо теплову карту з кореляцією (рис.17), (рис.18) матриця залежності (діаграми розсіювання).

Як бачимо на гістограмі та ядерної оцінці плотності розподілення є мультимодальне, тому типовим значенням - буде значення медіани (50%), вибірка невелика n = 25.

	curs	Agriculture	osvita	Med_15
count	25.000000	25.000000	25.000000	25.000000
mean	2598.480000	6511.160000	8743.320000	5722.200000
std	152.821006	2555.382403	2796.615793	2462.244708
min	2339.000000	2522.000000	4102.000000	2390.000000
25%	2526.000000	4265.000000	6215.000000	3467.000000
50%	2618.000000	6296.000000	8824.000000	5422.000000
75%	2731.000000	8730.000000	10995.000000	7081.000000
max	2827.000000	10508.000000	13034.000000	11346.000000

рис.14

```
[<AxesSubplot:title={'center':'osvita'}>,
 <AxesSubplot:title={'center':'Med_15'}>]],
```

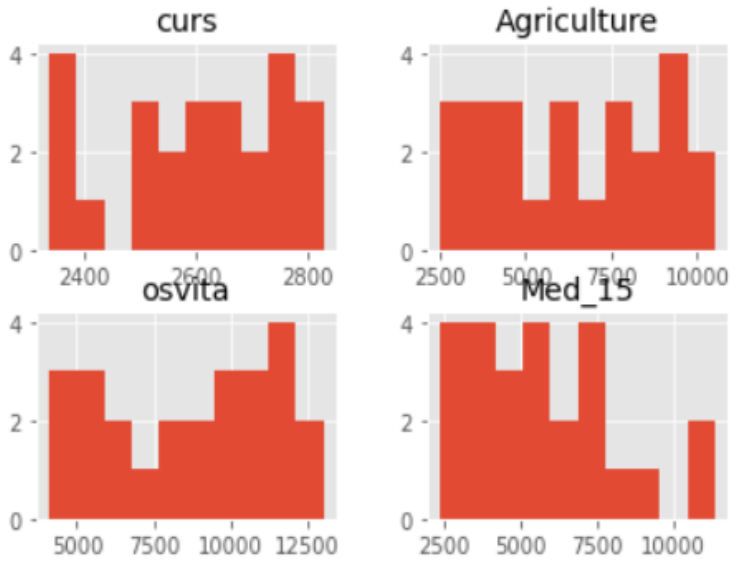


рис.15

```
#график
x = linspace(min(dt['osvita']), max(dt['osvita']), 1000)
plot(x, my_density(x), 'g')
```

```
]: [<matplotlib.lines.Line2D at 0x19ab1c86100>]
```

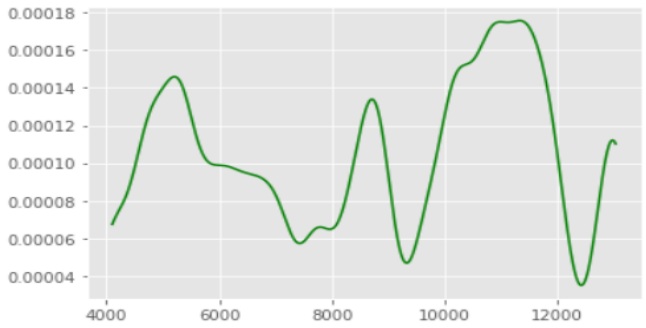


рис.16

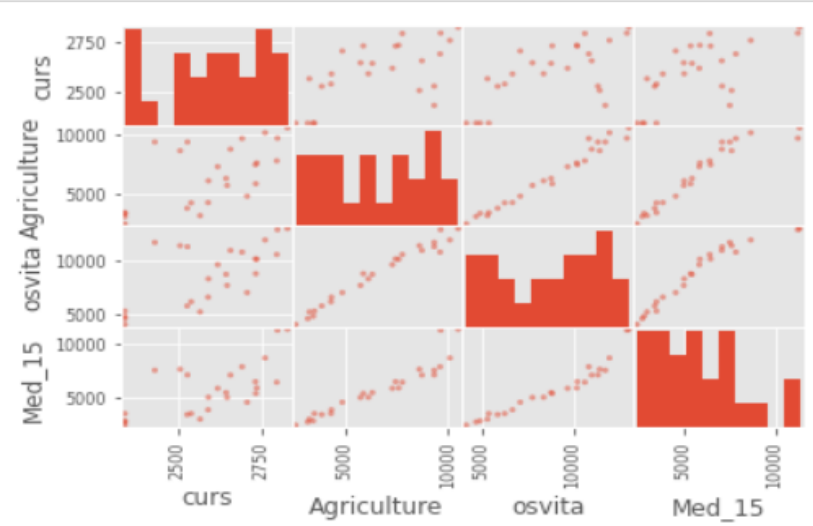
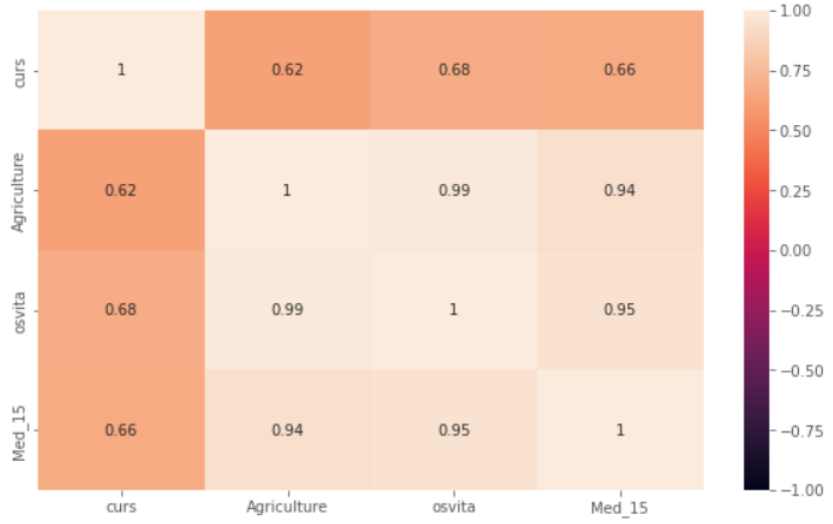

curs
Agriculture
osvita
Med_15
2750
2500
10000
5000
10000
5000
10000
5000
2500
2750
5000
10000
5000
10000
5000
10000
curs
Agriculture
osvita
Med_15

рис.17


curs
Agriculture
osvita
Med_15
1
0.62
0.68
0.66
0.62
1
0.99
0.94
0.68
0.99
1
0.95
0.66
0.94
0.95
1
curs
Agriculture
osvita
Med_15
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

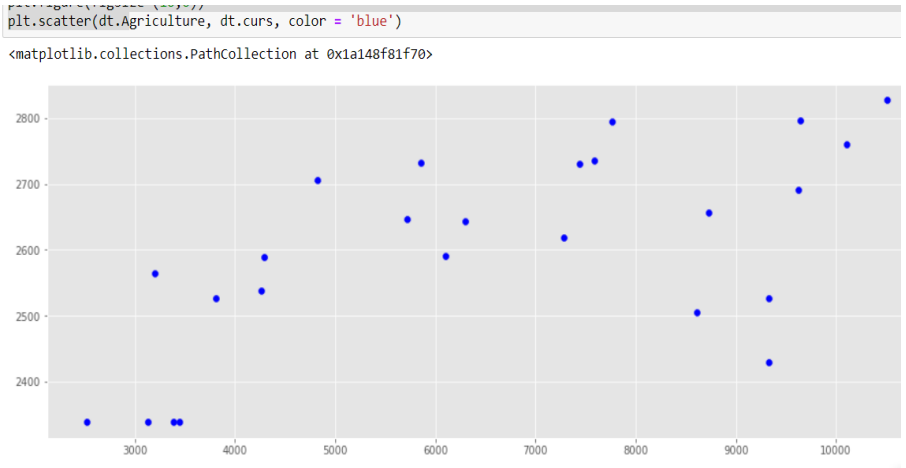

plt.scatter(dt.Agriculture, dt.curs, color = 'blue')
<matplotlib.collections.PathCollection at 0x1a148f81f70>
2800
2700
2600
2500
2400
3000
4000
5000
6000
7000
8000
9000
10000

рис.19

1. Для основного завдання використаємо такі ознаки:

- середній курс долара у квартал в період 2015-2021 р (незалежна величина)
- середні зарплати як залежні

2.1 Для вибірки для Лінійної та Поліноміальної регресії використовуємо ознаки:

- середній курс долара у квартал в період 2015- 2021 р (незалежна величина)
- середня заробітна плата в “сільському господарстві” в період 2015 - 2021р поквартально (залежна)

Для оцінки якості моделі використовуємо коефіцієнт детермінації R^2

2.2 Для Авторегресії та Аріми:

- період
- середня заробітна плата в “сільському господарстві” в період 2015 - 2021р поквартально (залежна)

Для оцінки якості моделі використаємо критерій Акаіке (AIC)

2.3 Підготовка даних для навчання та верифікації моделей

Щоб перевірити точність моделі, розділимо наші дані на навчальну та тестову вибірки (на тестову 30%). (рис.20) Використаємо навчальні дані для навчання нашої моделі, та перевіримо точність моделі на тестовій вибірці.

```
X = dt1.iloc[:, 2:3].values
y = dt1.iloc[:, -1].values
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
```

рис.20

Навчаємо нашу модель та будуємо графік (рис.21, рис.22)

```
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train, y_train)
```

рис.21

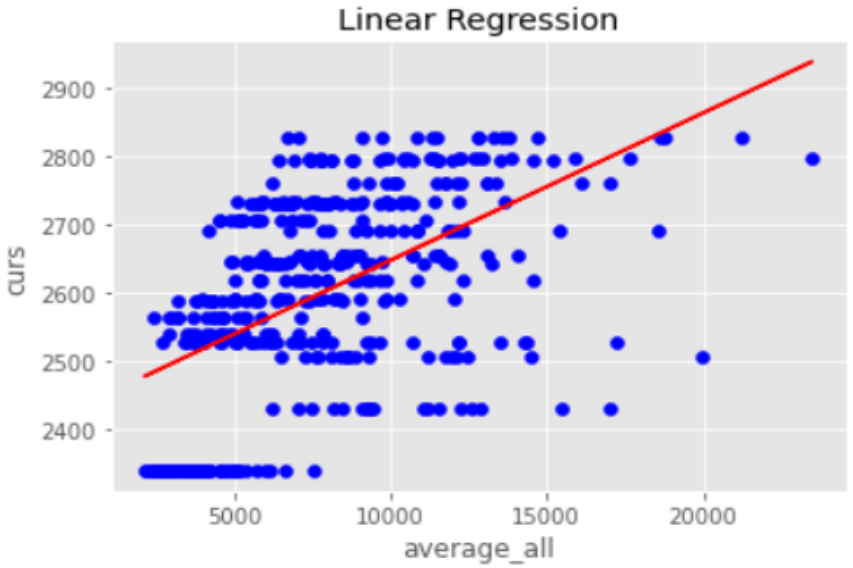


рис.22

Побудуємо Поліноміальну модель на тих самих даних, навчаємо нашу модель (рис.23), що використовували для Лінійної регресії (рис.24)

```
Polym = LinearRegression()
Polym.fit(X_polynom, y_train)
```

рис.23

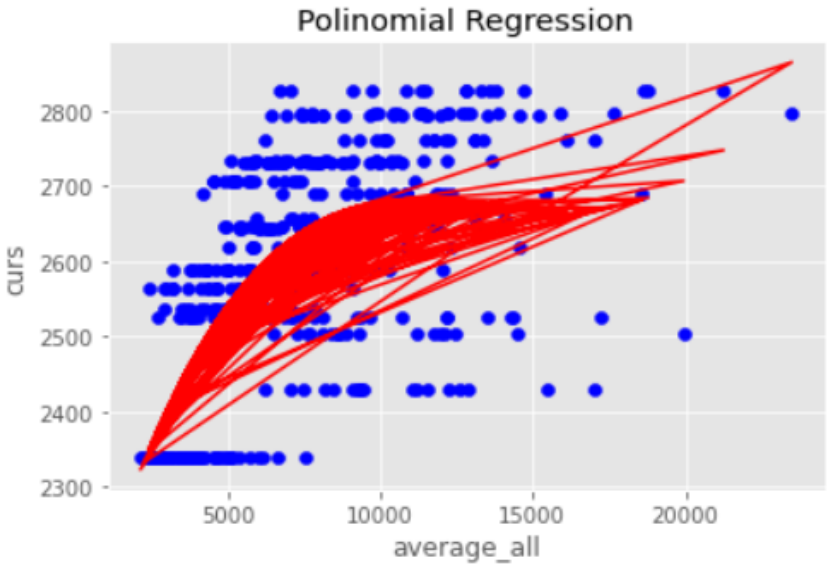


рис.24

Наша Лінійна та поліноміальна регресія для вибірки [2.1] (рис.25),(рис.26)

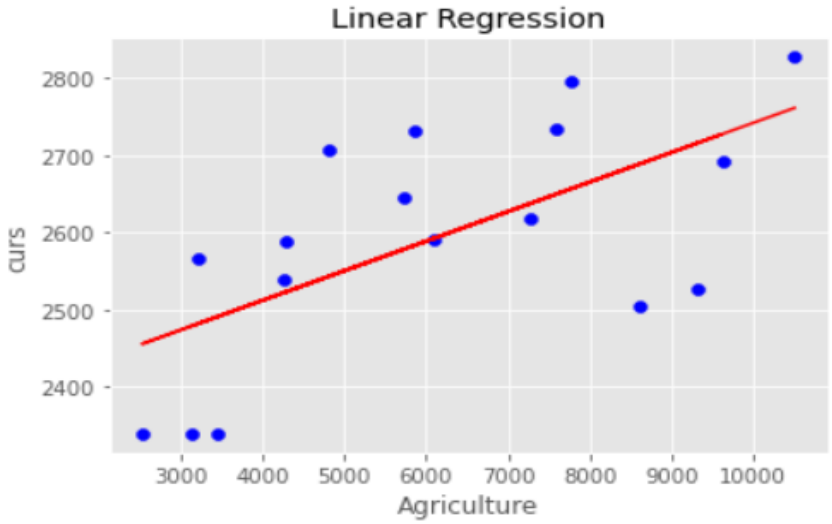


рис.25

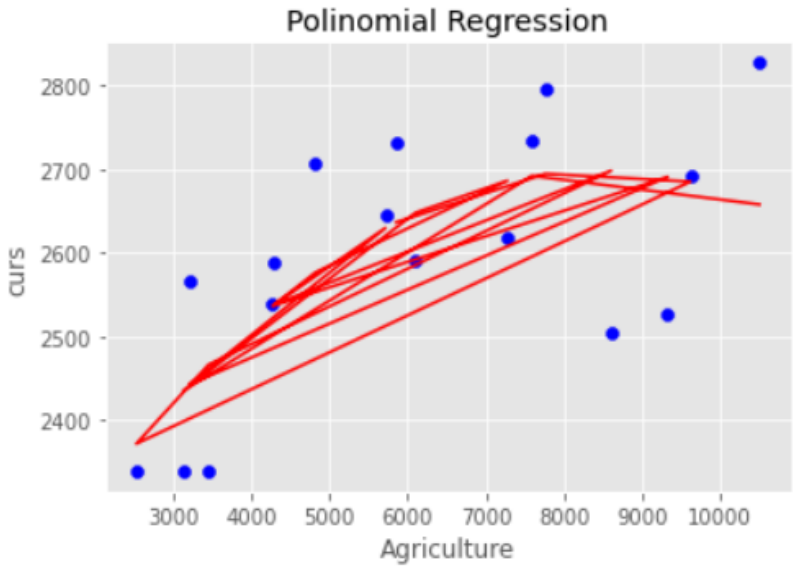


рис.26

Для Авторегресії побудуємо графік залежності заробітної плати в “сільському господарстві” від періоду (рис.27):

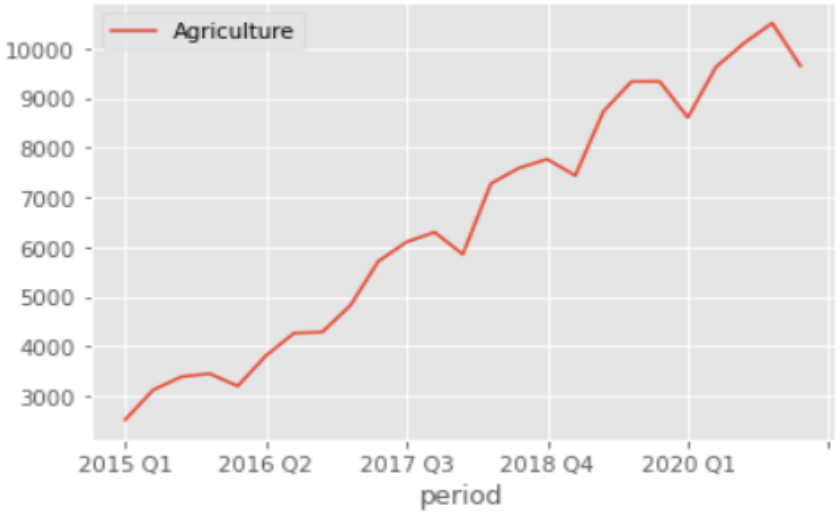


рис.27

Побудуємо графік з середнім та дисперсією (рис.28)

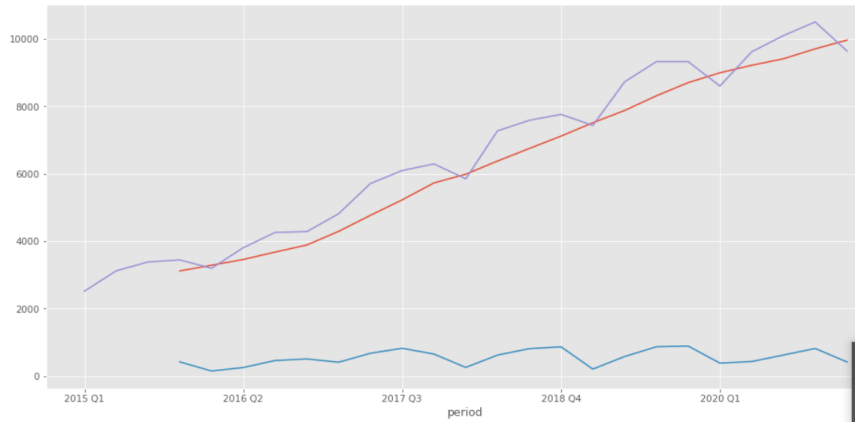


рис.28

Подивимося на значення, тренд, сезонність та залишки (рис.29)

рис.29

Тренд - лінійний не змінював свій характер до 3 квартала 2020 року, зростає.

Сезонність є, але це й не дивно, так як ми розглядаємо “сільське господарство”. Побудуємо графіки автокореляцій (рис.30)

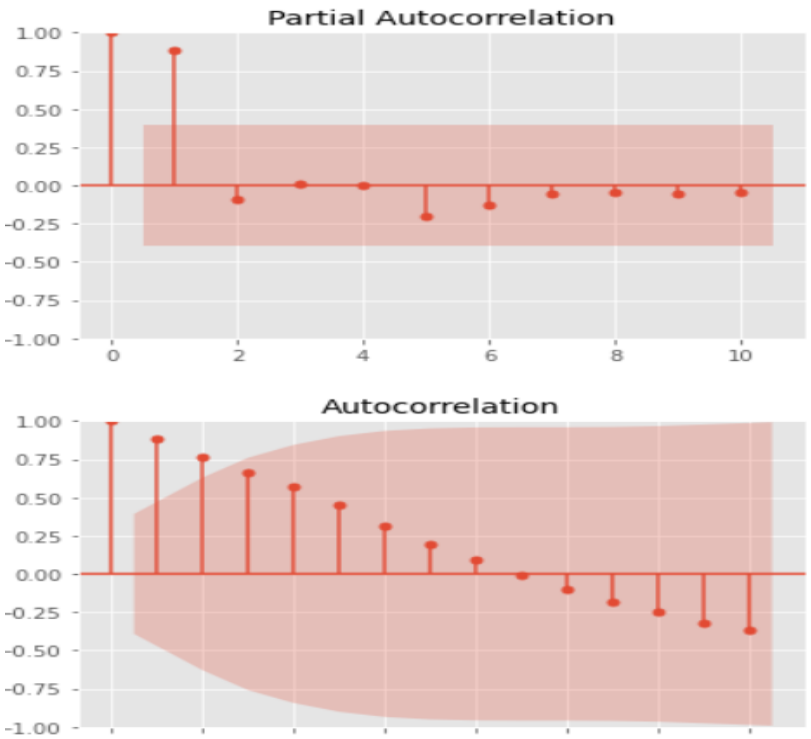


рис.30

В автокореляції бачимо повільне зменшення автокореляційної функції - це нам показує, що потрібно взяти 1 різницю, а в часній автокореляції маємо 1 пік - він в даний момент не так важливий. Візьмемо різницю, побудуємо графік (рис.31)

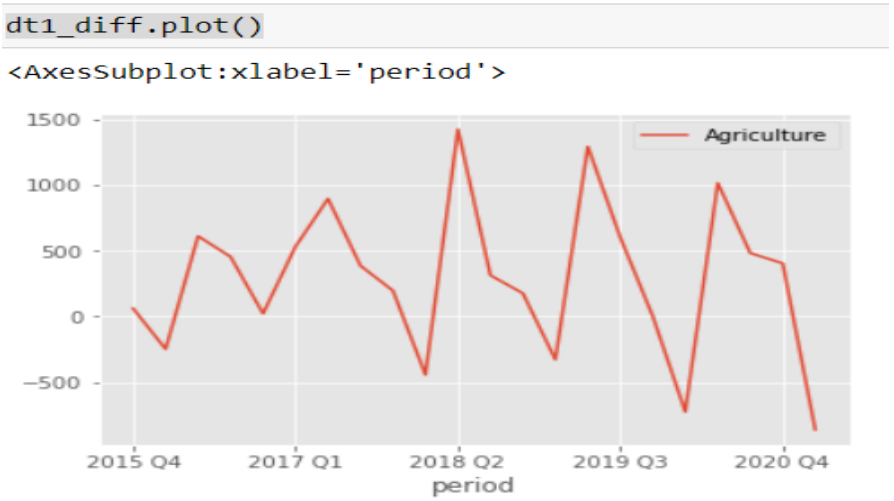


рис.31

Будуємо для ряду з різниць автокореляційну та часну автокореляційну функції (ри.32)

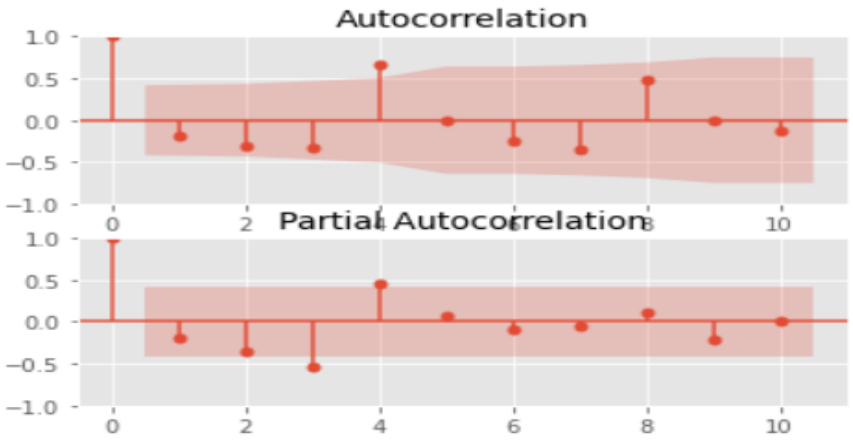


рис.32

Для прогноза краще використовувати останні значення ряда.

Знов розділяємо на навчальну та тестову вибірки (рис.33)

```
X = dt['Agriculture'].values
train = X[0:17] # 16 data as train data
test = X[16:] # 9 data as test data
predictions = []
```

рис.33

Навчаємо нашу модель на навчальній вибірці (рис.34)

```
from statsmodels.tsa.ar_model import AutoReg
from sklearn.metrics import mean_squared_error

model_ar = AutoReg(train, lags = 6)
model_ar_fit = model_ar.fit()
```

рис.34

Побудуємо модель Аріма та навчимо її на навчальній вибірці (рис.35)

```
from statsmodels.tsa.arima.model import ARIMA
```

```
model_arima = ARIMA(train,order=(1,1,1))
model_arima_fit = model_arima.fit()
```

2.4 Верифікація моделей

1. Результати моделей основного датасету, на тестових даних з оцінкою моделі з використанням R^2 (рис.35)- Лінійна регресія, (рис.36) - Поліноміальна регресійна модель:

```
y_predict_slr = lm.predict(X_test)

from sklearn import metrics
r_square = metrics.r2_score(y_test, y_predict_slr)
print('R-Square Error associated with Linear Regression: ', r_square)
```

```
R-Square Error associated with Linear Regression:  0.19277603596698
```

рис.35

```
y_predict_pr = Polym.predict(polynom.fit_transform(X_test))

from sklearn import metrics
r_square = metrics.r2_score(y_test, y_predict_pr)
print('R-Square Error associated with Polynomial Regression is: ', r_square)
```

```
R-Square Error associated with Polynomial Regression is:  0.30930697176933386
```

рис.36

2. Результати моделей для створеного датасету на базі основного, на тестових даних з оцінкою моделі з використанням R^2 (рис.37)- Лінійна регресія, (рис.38) - Поліноміальна регресійна модель та критерієм Акаіке (AIC) для Авторегресії (рис.39) та Аріми (рис.40):

```
y_predict_slr = lm.predict(X_test)

from sklearn import metrics
r_square = metrics.r2_score(y_test, y_predict_slr)
print('R-Square Error associated with Linear Regression: ', r_square)
```

```
R-Square Error associated with Linear Regression:  0.35840006862963014
```

рис.37

```
y_predict_pr = Polym.predict(polynom.fit_transform(X_test))

from sklearn import metrics
r_square = metrics.r2_score(y_test, y_predict_pr)
print('R-Square Error associated with Polynomial Regression is: ', r_square)
```

```
R-Square Error associated with Polynomial Regression is:  0.42734298546023464
```

рис.38

```
from statsmodels.tsa.ar_model import AutoReg
from sklearn.metrics import mean_squared_error

model_ar = AutoReg(train, lags = 6)
model_ar_fit = model_ar.fit()
print(model_ar_fit.aic)
```

```
174.44316692159873
```

рис.39

```
model_arima = ARIMA(train,order=(1,1,1))
model_arima_fit = model_arima.fit()
print(model_arima_fit.summary())
```

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   17
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -121.673
Date:                Fri, 24 Dec 2021   AIC                            249.347
Time:                        15:35:43   BIC                            251.665
Sample:                             0   HQIC                           249.466
                                  - 17
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9998      0.012     83.016      0.000       0.976       1.023
ma.L1         -0.9894      0.383     -2.581      0.010      -1.741      -0.238
sigma2      2.191e+05   1.78e-06   1.23e+11      0.000    2.19e+05    2.19e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.62   Jarque-Bera (JB):                 0.56
Prob(Q):                              0.43   Prob(JB):                         0.76
Heteroskedasticity (H):               3.93   Skew:                             0.45
Prob(H) (two-sided):                  0.16   Kurtosis:                         3.12
===================================================================================
```

рис.40

Будуємо графіки для Авторегресії (рис.41) та Аріми (рис.42) та графік наших даних (рис.43):

```
plt.plot(test)
plt.plot(predictions, color = 'blue')
```

```
[<matplotlib.lines.Line2D at 0x1a147d323d0>]
```

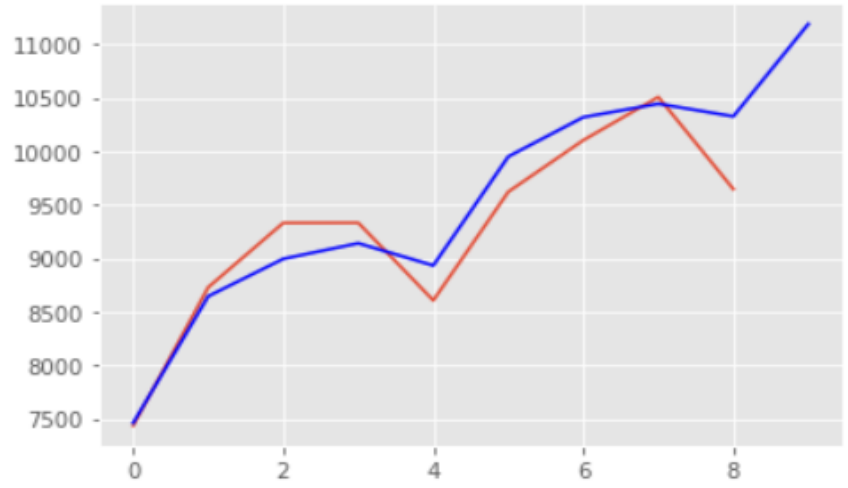


рис.41

```
plt.plot(test)
plt.plot(predictions, color = 'blue')
```

```
[<matplotlib.lines.Line2D at 0x1a14ae5e160>]
```

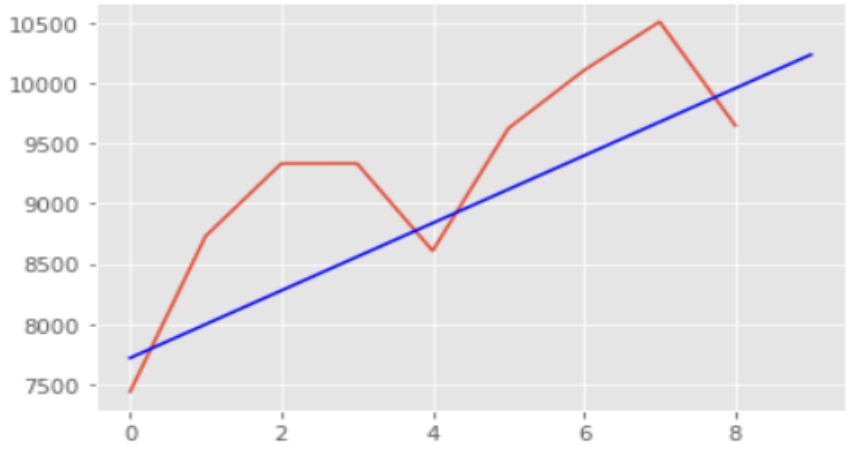


рис.42

```
dt['Agriculture'].plot()
```

```
<AxesSubplot:>
```

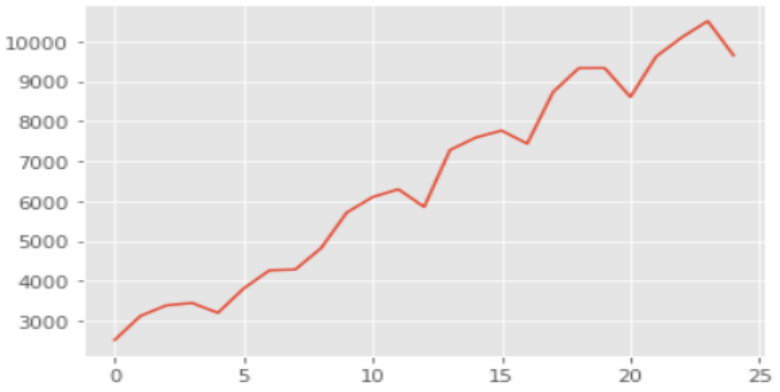


рис.43

Побудуємо автокореляційну та часну автокореляційну функції для залишків (рис.44):

```
residuals = model_arima_fit.resid
```

```
plt.figure()
plt.subplot(211)
plot_acf(residuals, lags=7, ax = plt.gca())
plt.subplot(212)

plot_pacf(residuals, method='ywmle', lags=7, ax = plt.gca()
plt.show()
```

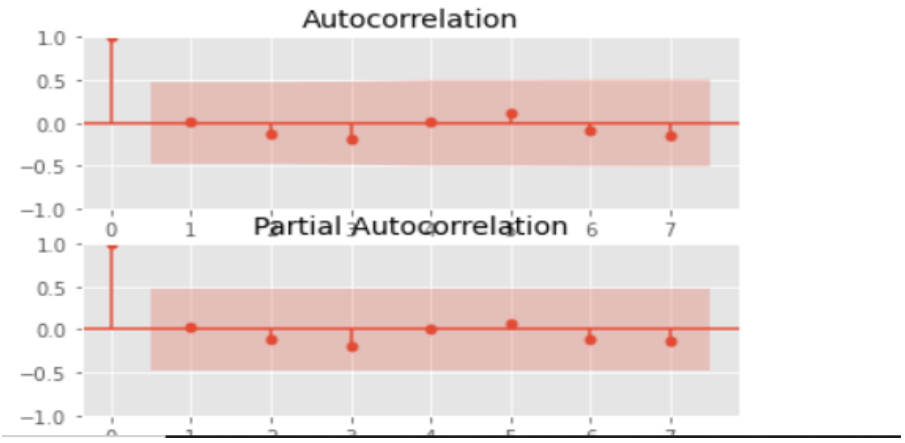


рис.44

2.5 Висновки щодо якості побудованих моделей

Для основного датасету обираючи між Лінійною і Поліноміальною регресією, можна віддати перевагу Лінійній, як простій моделі, легкій в коригування та інтерпретації, хоча вона показала гірший результат, коефіцієнт детермінації - 0.193, а Поліноміальної коефіцієнт детермінації - 0,309. Залежність між курсом долара та середньою заробітною платою є, але не велика та нелінійна, що показала нам кореляція, яка склала - 35%. Що є зрозумілим, так як ми розглядали середні заробітні плати у всіх галузях. Для перевірки, зробили вибірку заробітних плат з трьох

галузей, де кореляція вже склала > 62% , де Лінійна регресія показала непоганий результат - коефіцієнт детермінації 0,36, хоча також гірший від Поліноміальної регресії - 0,43. Для часових рядів краще впоралася Авторегресія, що видно на графіку, перевагу можна віддати їй в даний момент. Але для прогнозування на наступні квартали, цих даних для навчання може бути мало, потрібно врахувати вплив інших факторів, які вплинули на зміну характеру тренду в 3 кварталі 2020 року.

3. Результати аналізу

1. Заробітна плата залежить від курсу долара, але не тісно і не у всіх галузях.

2. Середня заробітна плата також залежить від статі, але щоб вказати, як сильно потрібно більше даних, більше інформації (такої як кількість працюючих жінок та чоловіків на однакових посадах у відповідних галузях).

3. Сезонність є, про що свідчать графіки, вплив має на залежні галузі, наприклад , як “сільське господарство”. Середня заробітна плата в Україні зростає з часом

Список використаних джерел

1. Статистика (модульний варіант з програмованою формою контролю знань). / Навчальний посібник/ А. Т. Опря / “Центр учбової літератури.” Київ 2012
2. [Електронний ресурс] Доступ до ресурсу: Алгоритмы машинного обучения для начинающих с примерами кода в Python
3. [Електронний ресурс] Доступ до ресурсу: Элбон Машинное обучение с использованием Python 2019
4. Закон України “Про оплату праці” [Електронний ресурс] Доступ до ресурсу: Про оплату праці | від 24.03.1995 № 108/95-ВР (rada.gov.ua)
5. [Електронний ресурс] Доступ до ресурсу: 5.1 Линейная модель | Прогнозирование: принципы и практика (2-е изд.) (otexts.com)
6. [Електронний ресурс] Доступ до ресурсу: Заробітна плата — Вікіпедія (wikipedia.org)