ДИПЛОМНА РОБОТА НА ТЕМУ: **«ТЕКСТОВИЙ АНАЛІЗ ДАНИХ ДЕКЛАРАЦІЙ НА ПРЕДМЕТ ВИЯВЛЕННЯ КОРУПЦІЇ»**

ВИКОНАВЕЦЬ РОБОТИ:

СТУДЕНТ IV КУРСУ,

ГРУПИ КА-53

ЯКУБЕЦЬ АНДРІЙ ОЛЕКСАНДРОВИЧ

КЕРІВНИК:

К.Т.Н., ДОЦЕНТ

КАФ. ММСА, ТЕРЕНТЬЄВ

ОЛЕКСАНДР МИКОЛАЙОВИЧ .

Київ-2019

МЕТА РОБОТИ

- СТВОРЕННЯ ПРОГРАМИ ДЛЯ ОБРОБКИ ВЕЛИКОЇ КІЛЬКОСТІ ДЕКЛАРАЦІЙ, ЩО ПРИШВИДШИТЬ РОБОТУ АНТИКОРУПЦІЙНИМ АГЕНТСТВАМ ДЛЯ ПОШУКУ КОРУПЦІЙНИХ СХЕМ ТА ПРОГНОЗУВАННЯ РІВНЯ КОРУПЦІЇ В УКРАЇНІ.

ОБ’ЄКТ ДОСЛІДЖЕННЯ

- ЩОРІЧНІ ДЕКЛАРАЦІЇ ЗА 2017 РІК РОЗМІЩЕНІ У ВІДКРИТОМУ ДОСТУПІ НА САЙТІ НАЦІОНАЛЬНОГО АГЕНТСТВА С ПИТАНЬ ЗАПОБІГАННЯ КОРУПЦІЇ.

МЕТОД ДОСЛІДЖЕННЯ

- ПАРСИНГ ДАНИХ, РОЗГЛЯД ТА АНАЛІЗ МЕТОДІВ РЕГРЕСІЙНОГО АНАЛІЗУ ТА ДИСПЕРСІЙНИЙ АНАЛІЗ.

АКТУАЛЬНІСТЬ РОБОТИ

- В УКРАЇНІ ПОСТІЙНО ВІДБУВАЮТЬСЯ РЕФОРМИ У ВСІХ СФЕРАХ ДЕРЖАВНОГО АПАРАТУ. ОДНОЮ ІЗ ТАКИХ РЕФОРМ – Є АНТИКОРУПЦІЙНА РЕФОРМА. СТВОРЕННЯ СИСТЕМИ, ЯКА ЗМОЖЕ ВИЯВЛЯТИ ДЕРЖАВНИХ СЛУЖБОВЦІВ ЗА ПРАВИЛАМИ РИЗИКУ ДОПОМОЖЕ ВИКОРІНИТИ КОРУПЦІЮ З НАЙВИЩИХ ЛАНОК.

- ЗА ЗАКОНОМ ПРО ЗАПОБІГАННЯ КОРУПЦІЇ — ВСІ ДЕРЖАВНІ СЛУЖБОВЦІ ЗОБОВ’ЯЗАНІ ЗАПОВНИТИ ТА ОПРИЛЮДНИТИ ДЕКЛАРАЦІЇ МАЙНОВОГО СТАНУ ДЛЯ ВІЛЬНОГО ДОСТУПУ ТА ЇХ ПЕРЕВІРКИ. ПОЧИНАЮЧИ З 2015 РОКУ, НА САЙТІ НАЦІОНАЛЬНОГО АГЕНТСТВА С ПИТАНЬ ЗАПОБІГАННЯ КОРУПЦІЇ З’ЯВЛЯЄТЬСЯ БІЛЬШЕ 1 МЛН. ДЕКЛАРАЦІЙ.

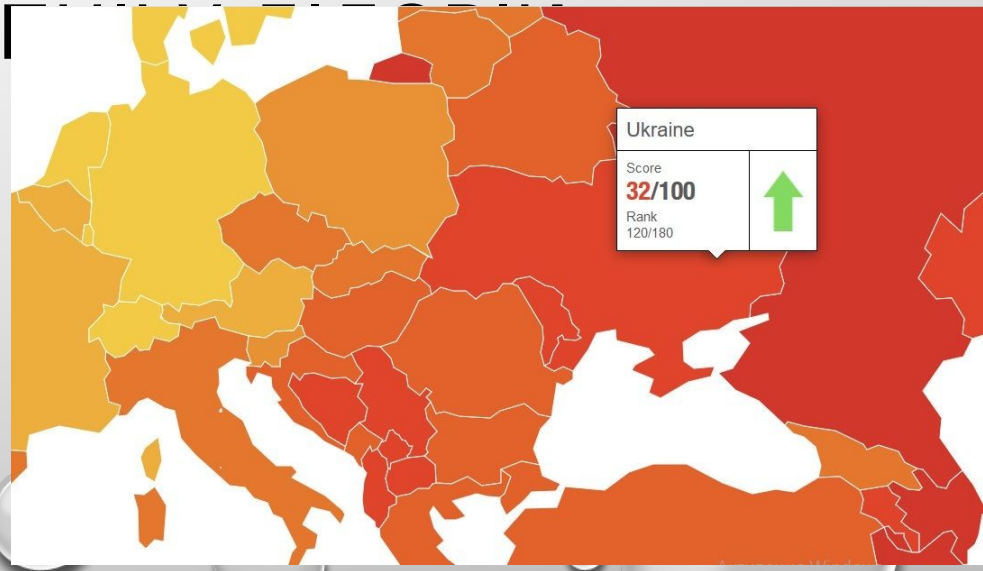
ПОСТАНОВКА ЗАДАЧІ

- РОЗРОБИТИ ПРОГРАМУ ДЛЯ ПАРСИНГУ, ОБРОБКИ ТА ЗАВАНТАЖЕННЯ ДАНИХ У ІНФОРМАЦІЙНО-АНАЛІТИЧНУ СИСТЕМУ;
- РОЗРОБИТИ СТРУКТУРИ ТАБЛИЦІ ЗНАЧЕНЬ РИЗИКІВ ДЕКЛАРАНТА, ЗА ЗАДАНИМИ ПРАВИЛАМИ РИЗИКУ ТА КРИТЕРІЯМИ ВІДБОРУ;
- РОЗРОБИТИ ПРОГРАМУ ДЛЯ ПРОГНОЗУВАННЯ КОРУПЦІЇ В УКРАЇНІ ТА ПРОВЕСТИ АНАЛІЗ НА СТАТИСТИЧНУ ЗНАЧИМІСТЬ РЕЗУЛЬТАТІВ ПРОГНОЗУ;
- ПРОТЕСТУВАТИ ПРОГРАМУ НА ЗАВАНТАЖЕНИХ ІЗ САЙТУ РЕАЛЬНИХ ДАНИХ.

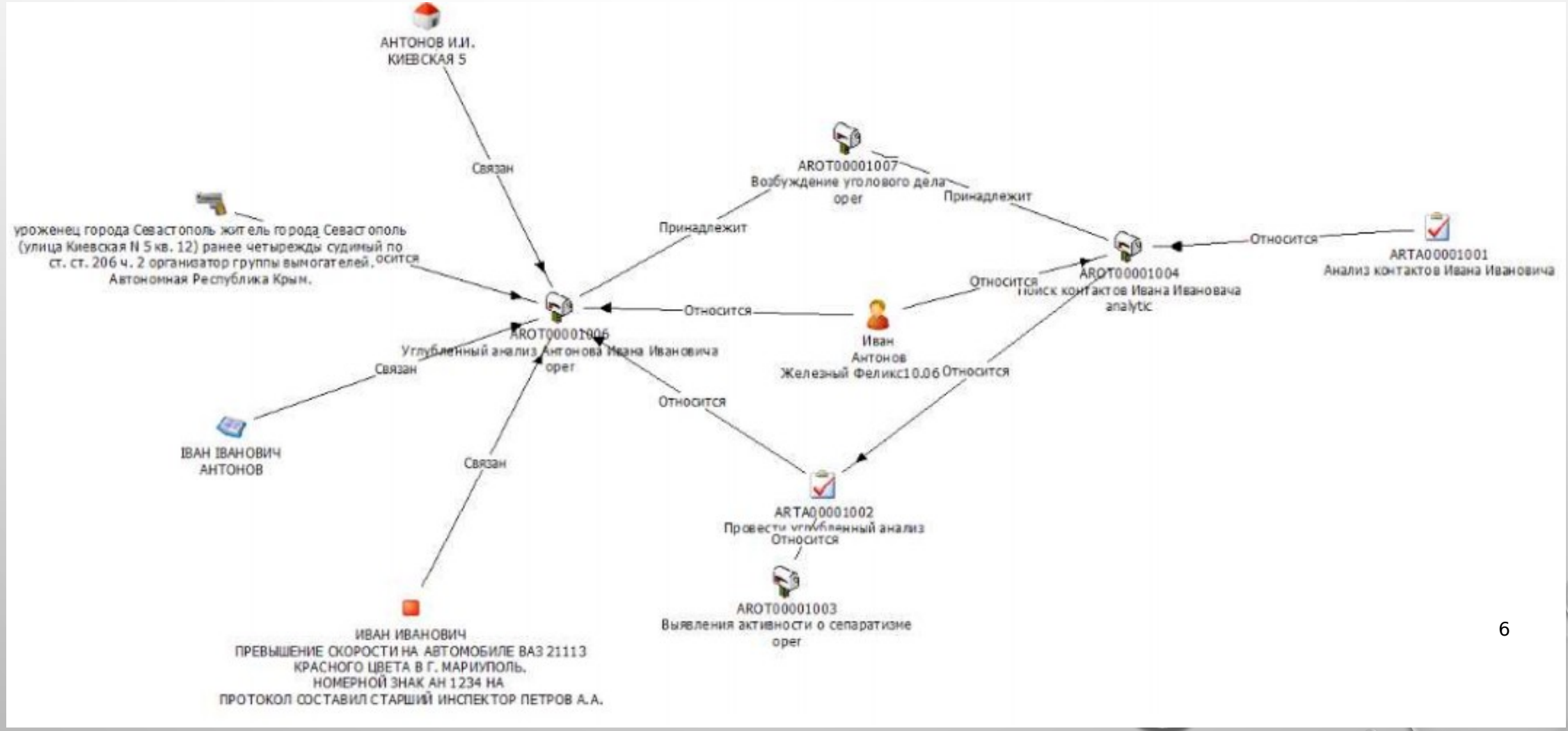
СТАН І ТЕНДЕНЦІЇ КОРУПЦІЇ В УКРАЇНІ

- У РЕЙТИНГУ ТІ УКРАЇНА СТАЛА НАЙКОРУМПОВАНІШОЮ КРАЇНОЮ ЄВРОПИ. У ДОСЛІДЖЕННІ ERNST&YOUNG ЗА 2017 РІК УКРАЇНА ЗАЙНЯЛА ПЕРШЕ МІСЦЕ ЗА ПОШИРЕНІСТЮ ХАБАРНИЦТВА/КОРУ
ПРАКТИЦІ

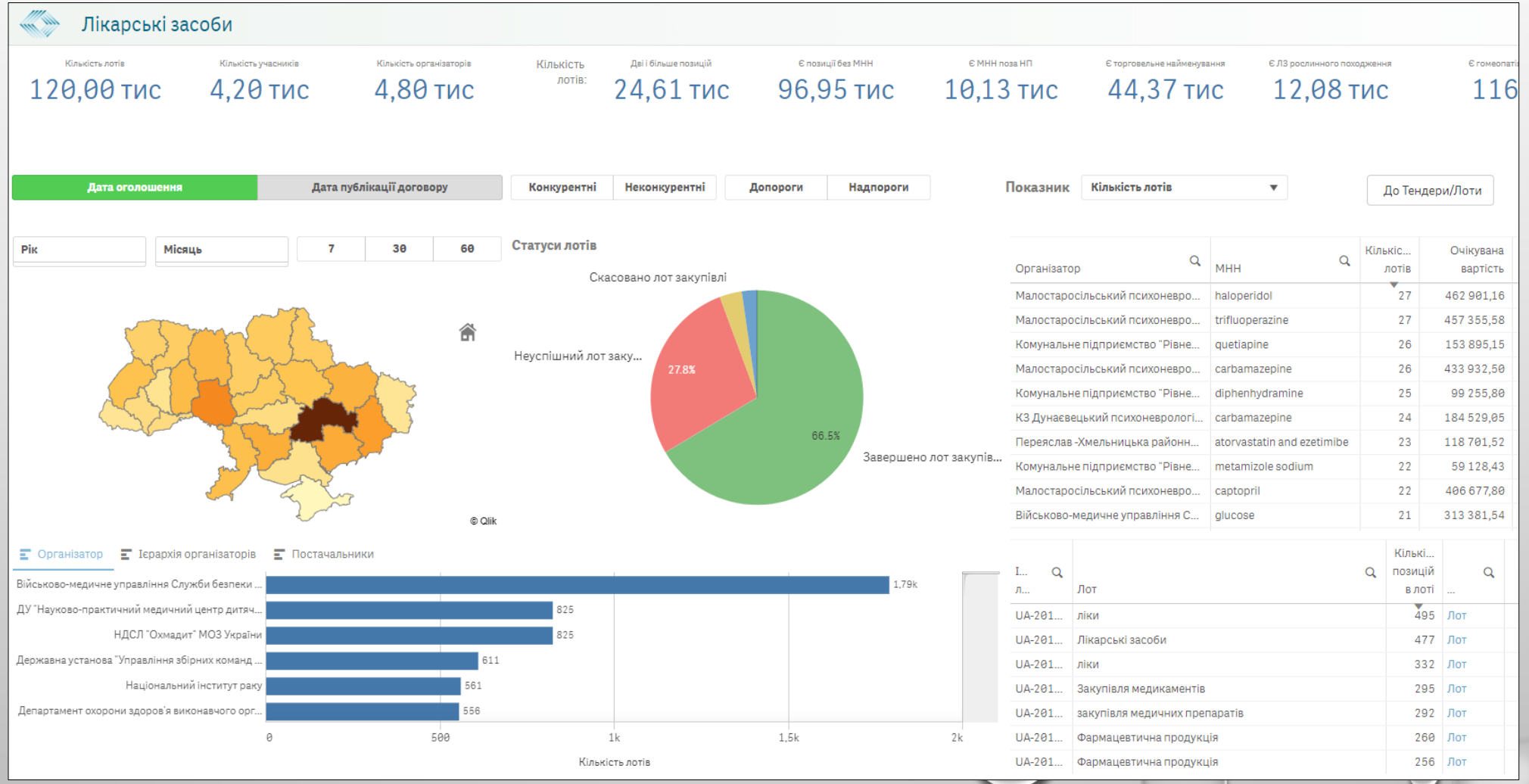
TRANSPARENCY INTERNATIONAL
the global coalition against corruption



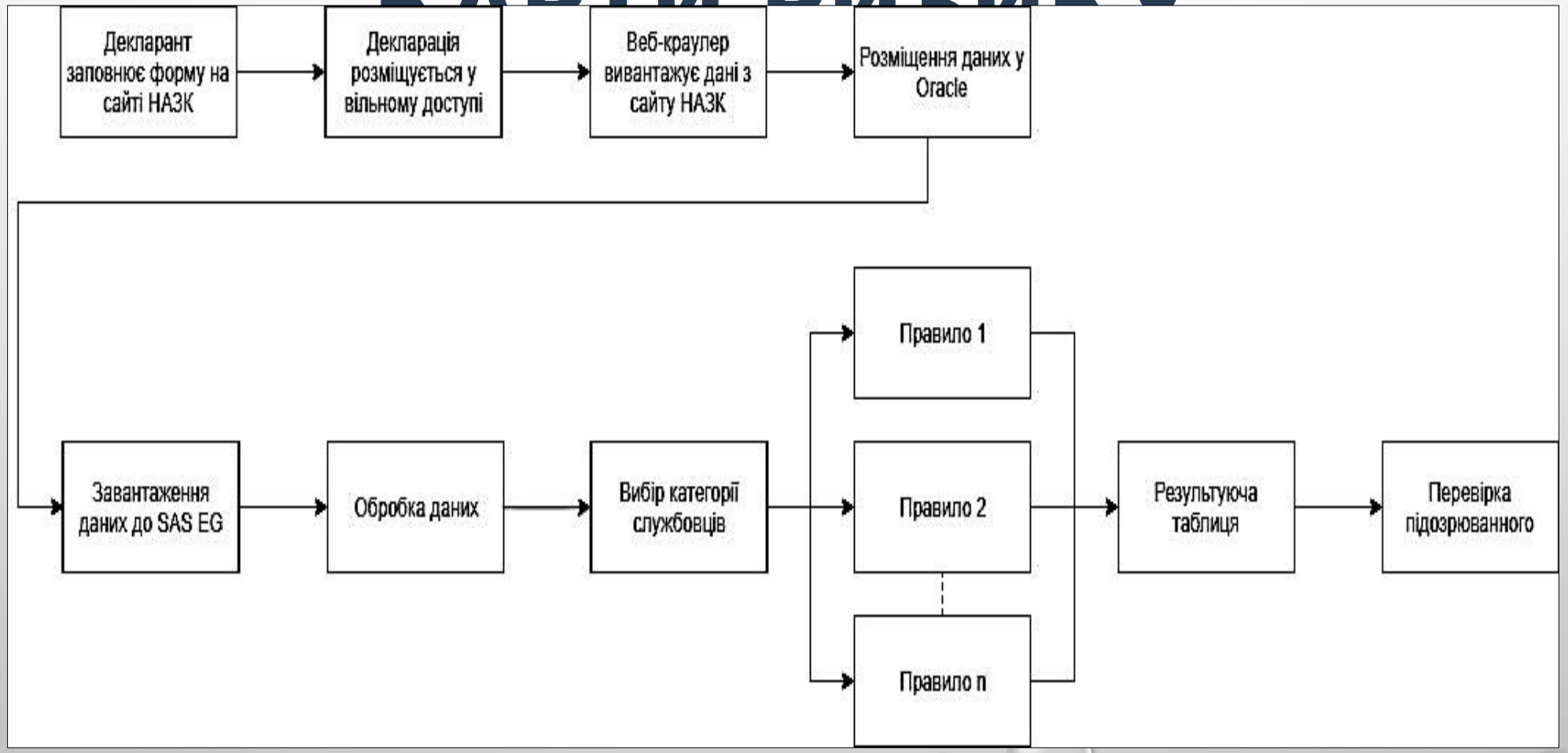
ОГЛЯД SAS® MEMEX



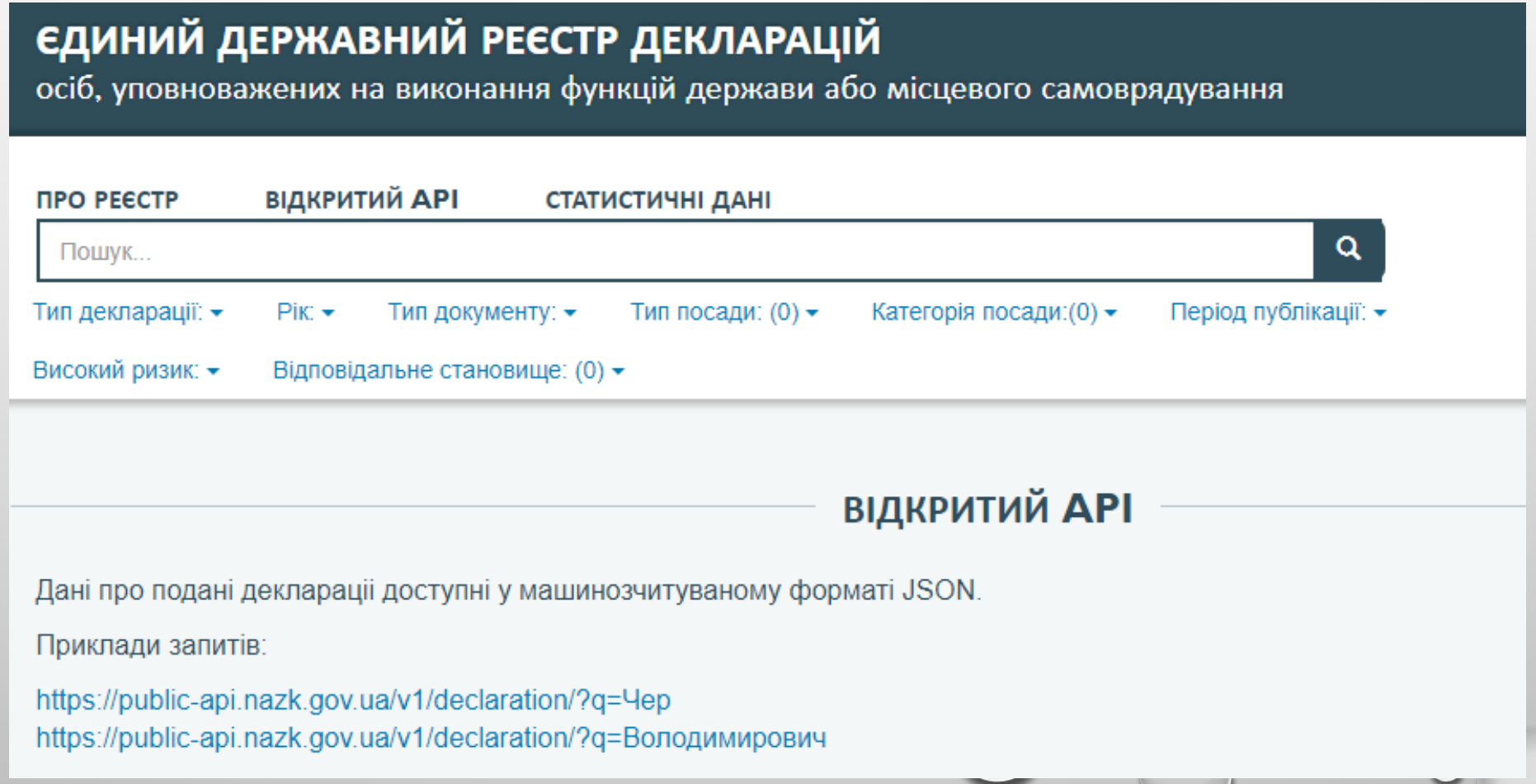
ОГЛЯД QLIK® SENSE



АРХІТЕКТУРА ПОБУДОВИ КАРТИ РИЗИКУ



ПАРСИНГ ДАНИХ З САЙТУ НАЗК



ВІДПОВІДЬ НА ЗАПИТ ПАРСЕРА

[{ "ID":"B238B773-316F-49E4-84A5-AEA634E18B65",
"CREATED_DATE":"05.09.2016",
"LASTMODIFIED_DATE":"05.09.2016",
"DATA":{
"STEP_0":{
"DECLARATIONTYPE":"1",
"DECLARATIONYEAR1":"2015" ...}]

КЛАСТЕРИЗАЦІЯ ДЕКЛАРАНТІВ ЗА ПОСАДОЮ

Term	Role	Attribute	Status	Weight	Imported Frequency
відділу	Noun	Alpha	Keep	0.128	12339
спеціаліст	Noun	Alpha	Keep	0.163	8395
+ начальник	Noun	Alpha	Keep	0.173	8006
головний	Noun	Alpha	Keep	0.194	6028
з	Noun	Alpha	Keep	0.219	5178
головний	Prop	Alpha	Keep	0.224	4328
управління	Noun	Alpha	Keep	0.234	4380
інспектор	Noun	Alpha	Keep	0.238	3743
+ депутат	Noun	Alpha	Keep	0.247	3382
+ заступник	Noun	Alpha	Keep	0.248	3398
державний	Noun	Alpha	Keep	0.268	2716
області	Noun	Alpha	Keep	0.274	2555
старший	Adj	Alpha	Keep	0.275	2514
+ сектор	Noun	Alpha	Keep	0.290	2131
сільської	Noun	Alpha	Keep	0.297	2019
питань	Noun	Alpha	Keep	0.299	2109
+ заступник начальник	Noun Group	Alpha	Keep	0.305	1823
забезпечення	Noun	Alpha	Keep	0.317	1776

РЕЗУЛЬТАТИ КЛАСТЕРИЗАЦІЇ

Cluster ID	Descriptive Terms	Frequency
4.0	відділення +командир служби військовослужбовець україни водій частини пожежний-рятувальник	4155.0
12.0	головний спеціаліст державний з інспектор питань призначення пенсій перерахунку осіб зборів п	3260.0
14.0	головний +бухгалтер спеціаліст державний ревізор-інспектор інспектор лікар +головной +'головно	3160.0
11.0	+начальник відділу +заступник +'заступник начальник' дільниці станції +'начальник караул' +караул	3125.0
7.0	секретар голова суддя +судовой засідання сільський сільської помічник +суд судді засідань судови	3018.0
1.0	депутат сільської селищної скликання працюю вчитель сільський ради підприємець приватний учи	2625.0
8.0	старший оперуповноважений слідчий інспектор державний офіцер майстер виконавець особливо	2220.0
5.0	поліції поліцейський інспектор патрульної +сектор реагування офіцер дільничний роти батальйон	2048.0
13.0	інспектор молодший відділу охорони і безпеки нагляду +режим категорії публічної інспектор-кінс	1348.0

РЕЗУЛЬТАТИ КЛАСТЕРИЗАЦІЇ

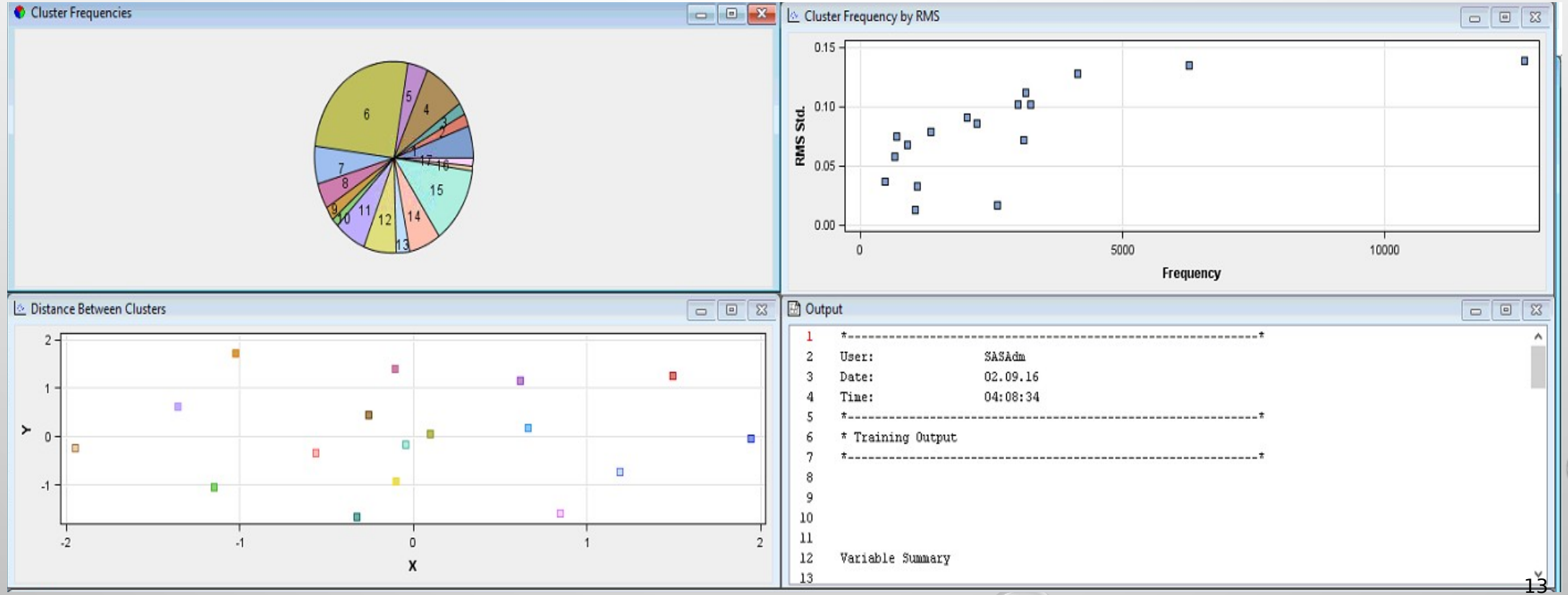
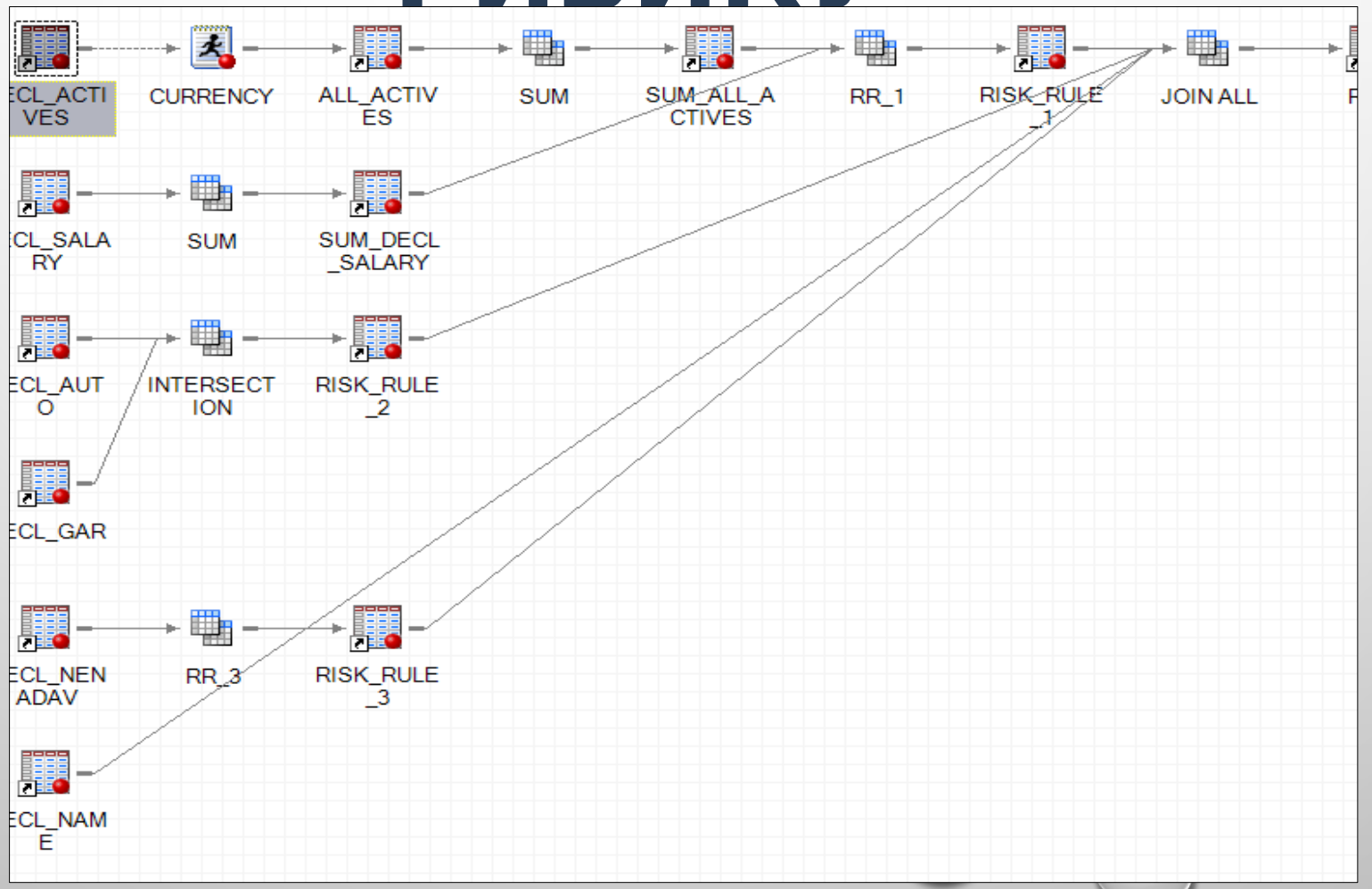


СХЕМА ПРОЦЕСУ КАРТИ РИЗИКУ



КАРТА РИЗИКУ ДЕКЛАРАЦІЙ

Изменить задачу | Фильтр и сортировка | Конструктор запросов | Where | Данные | Описательная статистика

ID	SUM_of_CURRENCY	SUM_of_SALARY	RATIO	RISK_RULE_
f8c6eb31-a33b-4272-8a6b-7e24ffc1b06e	315423	51	6184.7647059	
bab9be78-e50e-4bfe-b0e7-9b8a721e3643	808254	156	5181.1153846	
bb2066c5-faf7-4bad-8e0e-6b83830f778a	1717602	1159	1481.9689387	
b7fd7fde-87ad-45e8-979c-3ce4e89b6785	935481	766	1221.2545692	
bc84fd9f-411b-4749-92cd-729f47b1ff82	1828776	2938	622.45609258	
e84e51b7-0589-4564-90df-d2857357a733	1105702	2557	432.4215878	
e538fb0b-9763-4874-a589-f53e2fa5d432	1004934	2600	386.51307692	
e4aa1bad-1730-4963-8fc4-da296b2613f0	1546387	4237	364.97215011	
18a7db66-6dbf-4408-b8d0-12471588ccea	1528298	4897	312.08862569	
f13135a8-1402-429b-bf44-b539306161f1	223224	815	273.89447853	
e699fd5a-0352-4f22-8283-2e45a2bad02e	648426	2448	264.87990196	
b51e7878-29a4-413e-9724-1a903c5197d7	911038	3839	237.31127898	
ca8e2e60-40b1-42d8-8f22-ebc5d01eb12a	1182877	5638	209.80436325	
e97b91a8-f82d-4f42-9340-ebdc4d6e0c94	395691	1904	207.82090336	
174b5259-89dd-4e2d-ba4b-7c2b73b14830	1824901	9574	190.61008983	
e26ec8ad-7caa-4b73-9c2a-e7f229bfc4c2	649348	3916	165.81920327	
e26d1c57-2e7c-4044-903e-8f5520ed7ce8	1466991	8993	163.12587568	
abead4cd-fcb3-4412-b9df-b895261218bb	673302	4424	152.19303797	
f5b61d1d-576c-476f-abab-ba00036531bc	217423	1429	152.15045486	
b663d923-65a0-4dc0-9253-5954c1375ccb	983367	6587	149.2890542	
f347d133-8beb-43b6-bbff-ed2b88edb90d	394646	2719	145.14380287	
e07516cb-f4fa-46a7-b4af-08ccc21f25aa	888532	6671	133.1932244	
e373d65f-bae5-4152-a9b0-7bd6e19a1352	1146634	8660	132.40577367	
ce7998df-5a16-41c6-bbe7-953efeaecd64	749090	6342	118.11573636	
e303b7f4-d090-4073-9f03-05fee8e7ae4c	967357	8456	114.39888836	
b7fc7db4-2603-451c-bbf1-86b6553f1b1a	675755	6163	109.64708746	
e14d23bd-045a-46f8-b8c5-9d5e231c1d32	1480759	14290	103.62204339	
e50c3010-ad48-418d-9e72-6eb2e1d51ea2	953074	9539	99.913408114	

ПОБУДОВА ПРОГНОЗУЮЧОЇ МОДЕЛІ

Індекс Сприйняття корупції	CPI	Вимірює, наскільки поширена корупція в державному секторі даної країни. Країни оцінені від 0 до 10.	Рейтинг
ВВП	GDPpc	Вартість товарів і послуг, вироблених в одній країні поділений на населення даної країни.	Долари США
Рівень безробіття	UnempRate	Процент безробітних серед робочої сили.	Процент
Рівень		Темпи зростання споживчих	

РЕЗУЛЬТАТИ ПОБУДОВИ

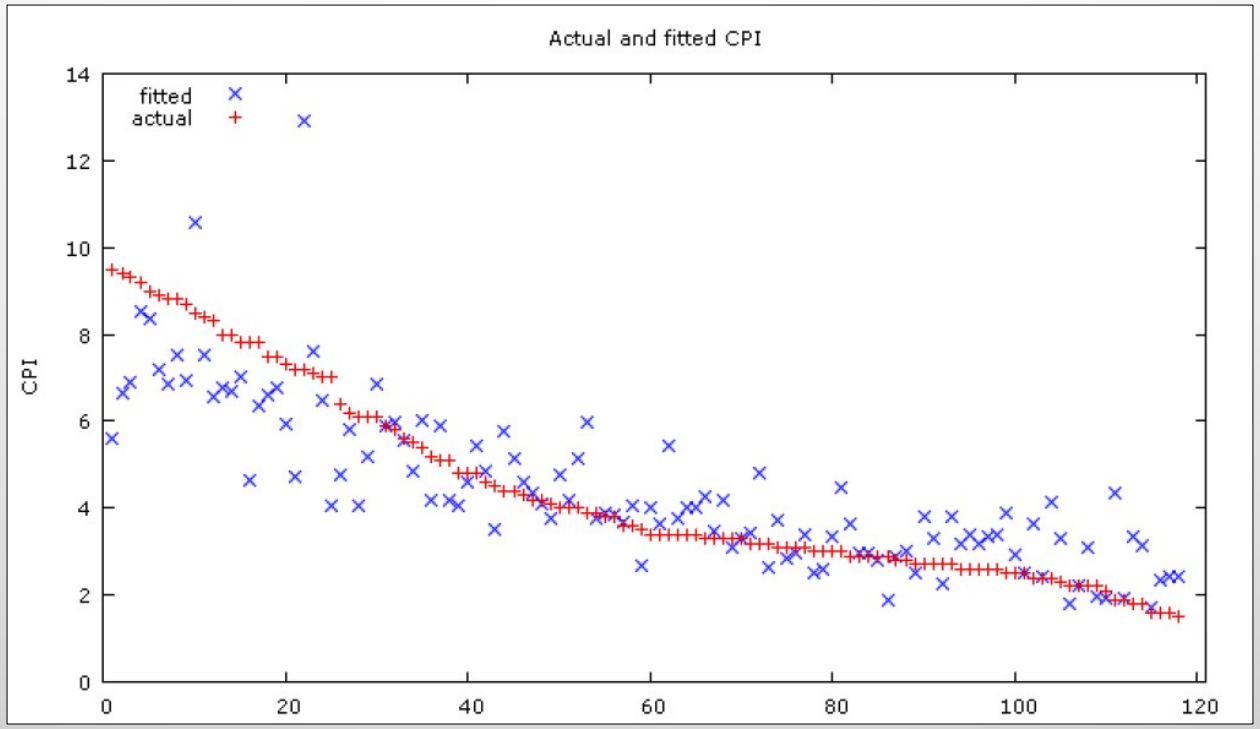
- $CPI = B_0 + B_1\,GDP - B_2\,UNEMPRATE - B_3 INFLARATE + U_i$

Source	SS	df	MS	Number of obs = 118
				F(3, 114) = 84.75
Model	405.197884	3	135.065961	Prob > F = 0.0000
Residual	181.676347	114	1.59365217	R-squared = 0.6904
				Adj R-squared = 0.6823
Total	586.874231	117	5.01601907	Root MSE = 1.2624

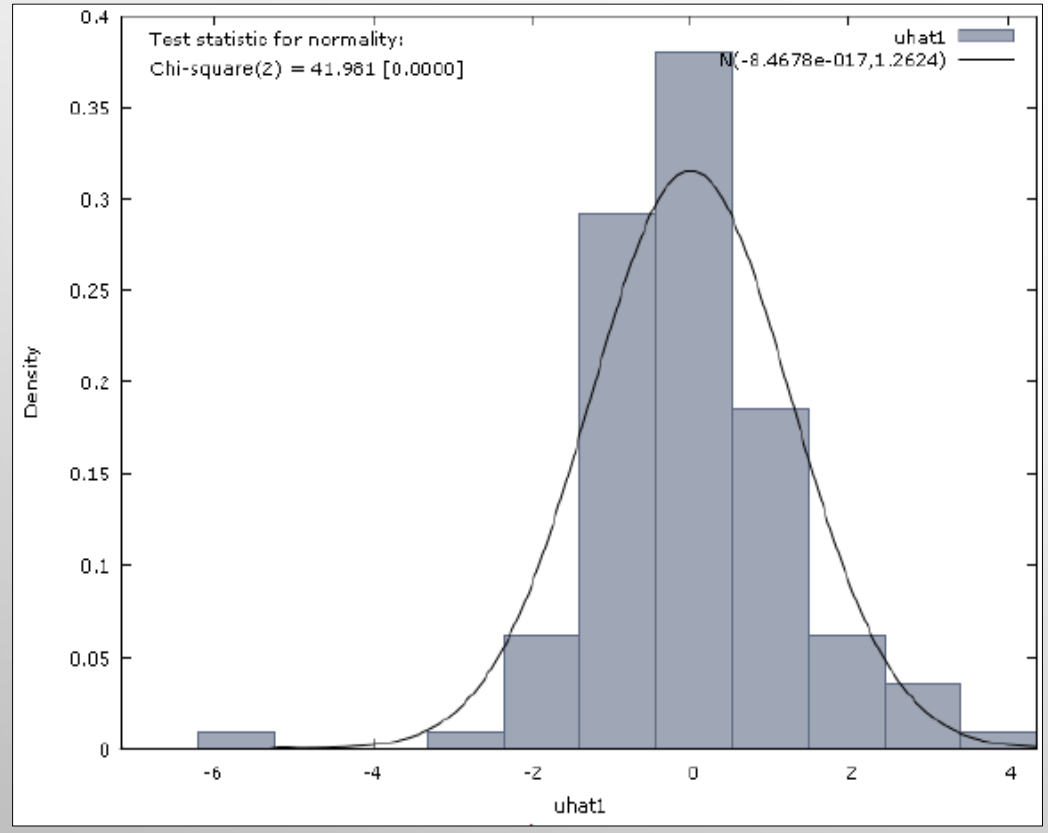
CPI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
GDPpc	.0000918	8.02e-06	11.46	0.000	.0000759	.0001077
UnempRate	-.0112513	.0082036	-1.37	0.173	-.0275025	.0049999
InflaRate	-.1030416	.030841	-3.34	0.001	-.1641375	-.0419458
_cons	3.533943	.3461715	10.21	0.000	2.84818	4.219706

Fig A[1]

ГРАФІК ЗАЛЕЖНОСТІ МІЖ РЕАЛЬНИМИ ЗНАЧЕННЯМ ТА ЗМОДЕЛЬОВАНИМИ



ОЦІНКА РЕЗУЛЬТАТІВ

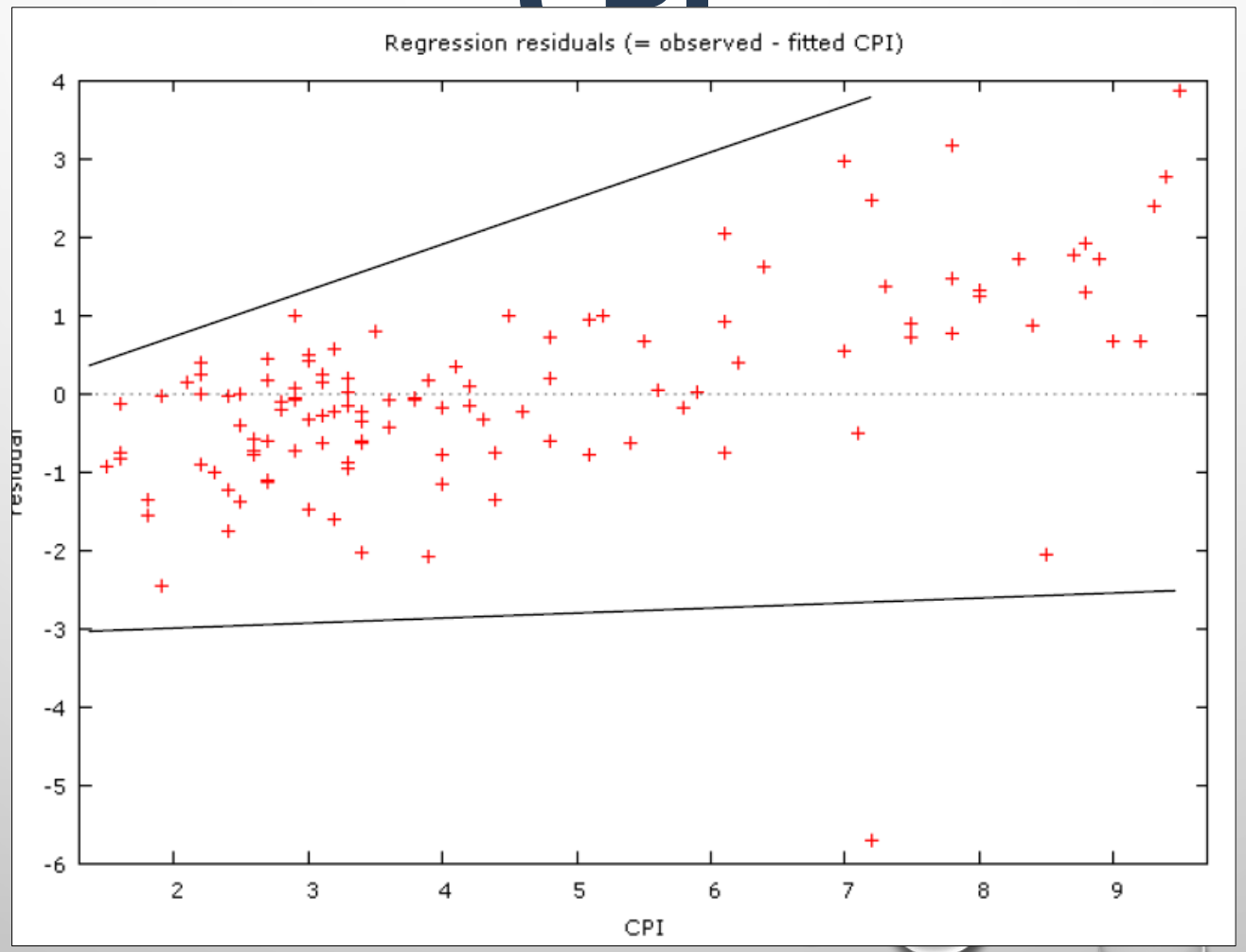


	Sum of squares	df	Mean square
Regression	405.198	3	135.066
Residual	181.676	114	1.59365
Total	586.874	117	5.01602

R^2 = 405.198 / 586.874 = 0.690434

F(3, 114) = 135.066 / 1.59365 = 84.7525 [p-value 6.72e-029]

ГРАФІК ЗАЛИШКІВ ВІДНОСНО CPI



ВИСНОВКИ

- РОЗРОБЛЕНО ПРОГРАМУ ДЛЯ ПАРСИНГУ, ОБРОБКИ ТА ЗАВАНТАЖЕННЯ ДАНИХ У ІНФОРМАЦІЙНО-АНАЛІТИЧНУ СИСТЕМУ;
- РОЗРОБЛЕНО СТРУКТУРУ ТАБЛИЦІ ЗНАЧЕНЬ РИЗИКІВ ДЕКЛАРАНТА, ЗА ЗАДАНИМИ ПРАВИЛАМИ РИЗИКУ ТА КРИТЕРІЯМИ ВІДБОРУ;
- РОЗРОБЛЕНО ПРОГРАМУ ДЛЯ ПРОГНОЗУВАННЯ КОРУПЦІЇ В УКРАЇНІ ТА ПРОВЕСТИ АНАЛІЗ НА СТАТИСТИЧНУ ЗНАЧИМІСТЬ РЕЗУЛЬТАТІВ ПРОГНОЗУ;
- ПРОТЕСТОВАНО ПРОГРАМУ НА ЗАВАНТАЖЕНИХ ІЗ САЙТУ РЕАЛЬНИХ ДАНИХ.

ПЕРСПЕКТИВИ ЩОДО ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

- СТВОРЕННЯ ЗВ’ЯЗКУ МІЖ ДЕКЛАРАЦІЯМИ ОДНОГО І ТОГО САМОГО ДЕКЛАРАНТА ЗА РІЗНІ РОКИ
- СТВОРЕННЯ ПЗ ДЛЯ АВТОМАТИЧНОГО ОНОВЛЕННЯ БАЗИ ДАНИХ ДЕКЛАРАЦІЙ
- ДОДАННЯ НОВИХ ПРАВИЛ ДЛЯ КАРТИ РИЗИКУ
- ВПРОВАДЖЕННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПОШУКУ СЛІДІВ КОРУПЦІЙНОГО ЗЛОЧИНУ

ПУБЛІКАЦІЇ ТА УЧАСТЬ У КОНФЕРЕНЦІЯХ

- ТЕРНЕТЬЄВ О.М., ЯКУБЕЦЬ А.О., ПРОСЯНКІНА-ЖАРОВА Т.І. **ЗАСТОСУВАННЯ SAS ENTERPRISE GUIDE ДЛЯ ВИЯВЛЕННЯ ЗВ’ЯЗКІВ У ПРЕДМЕТНО-ОРІЄНТОВАНИХ ДАНИХ** ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ: ТЕОРІЯ І ПРАКТИКА: ТЕЗИ ДОПОВІДЕЙ ІІ ВСЕУКРАЇНСЬКОЇ ІНТЕРНЕТ-КОНФЕРЕНЦІЇ ЗДОБУВАЧІВ ВИЩОЇ ОСВІТИ І МОЛОДИХ УЧЕНИХ (4 КВІТНЯ 2019 Р., М. ЗАПОРІЖЖЯ)

ДЯКУЮ ЗА УВАГУ!