Практическая работа

Анализ данных по США 1970-х для определения уровня социального развития

Целью работы является классификация по качеству жизни всех штатов по следующим признакам (критериям):

1) Высокий доход на душу населения;

2) Высокий уровень образования;

3) Благополучная криминальная обстановка;

4) Хорошая продолжительность жизни;

5) Благоприятный климат.

Исходный набор данных (датасет) находится в файле statedata.csv.

Для каждого штата набор данных включает такие поля данных:

- ❖ население,
- ❖ доход на душу населения,
- ❖ уровень неграмотности,
- ❖ уровень убийств,
- ❖ количество выпускников средней школы,
- ❖ среднее количество морозных дней,
- ❖ площадь суши,
- ❖ регион (экономический), к которому относится штат,
- ❖ аббревиатура (код) из двух букв.

В датасете есть 50 записей, по одной для каждого штата, содержащих 11 переменных:

- ✓ Population (население) – численность населения в штате в 1975 г. (в тыс.чел).
- ✓ Income (доход) – доход на душу населения в 1974 г. (за год)
- ✓ Illiteracy (неграмотность) – уровень неграмотности в 1970 г. (в процентах к населению).
- ✓ Life_Exp (продолжительность жизни) – ожидаемая продолжительность жизни в годах жителей штата в 1970 г.
- ✓ Murder (убийства) – количество убитых (как умышленно, так и случайно) на 100 000 населения в 1976 г.
- ✓ HS_Grad (выпускники) – процент выпускников средней школы в 1970 г.
- ✓ Frost (мороз) – среднее число дней с минимальной температурой ниже нуля с 1931–1960 гг. в столице штата.
- ✓ Area (площадь) – площадь суши (в квадратных милях) штата.
- ✓ State_name – полное название штата.
- ✓ State_abbr – аббревиатура (код) для каждого штата.
- ✓ State_division – регион, к которому принадлежит штат (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, Pacific).

Импортируйте датасет в новый лист файла Excel. Для чтения CSV используйте мастер чтения текстовых файлов с использованием разделителя полей «запятая» и с учетом разделителя в дробных числах «точка». Назовите лист в Excel “usa-stat”.

Сделайте из полученных данных «умную таблицу», для этого поставьте курсор внутри блока ячеек с данными и нажмите Ctrl-T (или через меню Главная – Стили – Форматировать как таблицу). Дайте имя этой таблице – Штаты.

Примените условное форматирование для выделения благоприятных и проблемных штатов по переменным murder, illiteracy, life_exp, income, HS_Grad, frost. (рисунок 1)

Будем использовать такие настройки правил условного форматирования:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Population	Income	Illiteracy	Life_Exp	Murder	HS_Grad	Frost	Area	State_name	Region	State_Abbr	плот_на
2	21198	5114	1,1	71,71	10,3	62,6	20	156361	California	Far West	CA	0,136
3	868	4963	1,9	73,6	6,2	61,9	0	6425	Hawaii	Far West	HI	0,135
4	3559	4864	0,6	71,72	4,3	63,5	32	66570	Washington	Far West	WA	0,053
5	2284	4660	0,6	72,13	4,2	60	44	96184	Oregon	Far West	OR	0,024
6	590	5149	0,5	69,03	11,5	65,2	188	109889	Nevada	Far West	NV	0,005

Рисунок 1

Подготовьте новый блок данных о регионах на основе «умной таблицы», для чего в контекстном меню включите Таблица -- Строка итогов. В полученной под таблицей данных строке подитогов установите для переменных murder, illiteracy, life_exp, income, frost, HS_Grad вариант подсчета – функцию Среднее, а для переменных Population, Area – функцию Сумма (рисунок 2).

45	3615	3624	2,1	69,05	15,1	41,3	20	50708	Alabama	Southeast	AL	0,071
46	2341	3098	2,4	68,09	12,5	41	50	47296	Mississippi	Southeast	MS	0,049
47	2110	3378	1,9	70,66	10,1	39,9	65	51945	Arkansas	Southeast	AR	0,041
48	12237	4188	2,2	70,9	12,2	47,4	35	262134	Texas	Southwest	TX	0,047
49	2715	3983	1,1	71,42	6,4	51,6	82	68782	Oklahoma	Southwest	OK	0,039
50	2212	4530	1,8	70,55	7,8	58,1	15	113417	Arizona	Southwest	AZ	0,020
51	1144	3601	2,2	70,32	9,7	55,2	120	121412	New Mexico	Southwest	NM	0,009
52	**212321**	**4415,8**	**1,17**	**70,8786**	**7,378**	**53,108**	**104,16**	**3536794**			**50**	**0,15**

Рисунок 2

Используя фильтр в заголовке таблицы отфильтруйте по очереди все регионы, и скопируйте получаемую строку итоговых значений в новую таблицу (начиная с ячейки B57). Затем в ячейках столбца A впишите названия соответствующих регионов, а в строке 56 названия переменных (столбцов): регион, доход, неграмотность, продолж_жизни, убийства, сред_образ, мороз, площадь, плот_нас.

Сделайте также из полученных данных «умную таблицу», для этого поставьте курсор внутри блока ячеек с данными и нажмите Ctrl-T (или через меню Главная – Стили – Форматировать как таблицу). Дайте имя этой таблице – Регионы (рисунок 3).

56	регион	доход	неграмотнос	продолж_жизни	убийств	сред_образ	мороз	площад	плот_нас
57	Far West	5010,83	1,03	71,25	7,97	63,32	74,33	1001861	0,06
58	Great Lakes	4669,00	0,80	70,99	7,78	53,20	129,40	244101	0,17
59	Mideast	4939,40	1,06	70,44	7,38	52,46	105,40	112191	0,47
60	New England	4423,83	0,92	71,58	3,38	55,05	144,00	62951	0,41
61	Plains	4569,71	0,63	72,32	3,49	55,46	145,00	507723	0,03
62	Rocky Mountain	4387,60	0,62	71,54	5,70	62,56	148,80	511329	0,01
63	Southeast	3826,00	1,89	69,39	11,33	41,97	59,42	530893	0,09
64	Southwest	4075,50	1,83	70,80	9,03	53,08	63,00	565745	0,03
65	Итог	4487,74	1,10	71,04	7,01	54,64	108,67	3536794	

Рисунок 3

Определите в итоговой строке средние значения по регионам – по группам штатов – по переменным murder, illiteracy, life_exp, income, frost, HS_Grad, и суммарные значения по переменным –Population, Area.

Добавьте в таблице Штаты новое поле для значения плотности населения с именем «Плот_нас», в котором введите формулу расчета = [Population]/[Area]. В строке поитогов установите функцию Среднее.

Проведите ранжирование регионов по основным показателям. Для этого скопируйте в столбик перечень названий регионов, например, с ячейки A68. Затем подготовьте столбцы для ранжирования: убийства, продолж-жизни, доход, образование, интеграл_рейтинг. В столбцах проведите ранжирование, соответственно,

в ячейке B68 формулой РАНГ(E57;Таблица2[убийства]),
в ячейке C68 формулой РАНГ(D57;Таблица2[продолж_жизни] ;1),
в ячейке D68 формулой РАНГ(B57;Таблица2[доход];1),
в ячейке E68 формулой РАНГ(F57; Таблица2[сред_образ];1).

Затем для интегрального рейтинга примем формулу взвешенной суммы рейтингов, т.е. для F68 введем формулу =СУММ(B68:E68)/4.

Примените условное форматирование для результатов «интеграл_рейтинг» с применением значков (рисунок 4).

67		убийства	продолж-жизни	доход	образован	интеграл.рейтинг
68	Far West	3	5	8	8	6,00
69	Great Lakes	4	4	6	4	4,50
70	Mideast	5	2	7	2	4,00
71	New England	8	7	4	5	6,00
72	Plains	7	8	5	6	6,50
73	Rocky Mountain	6	6	3	7	5,50
74	Southeast	1	1	1	1	1,00
75	Southwest	2	3	2	3	2,50

Рисунок 4

Найдите топ-3 штатов (с помощью сортировки по соответствующему полю по убыванию/возрастанию) для следующих показателей:

- самые густонаселенные и малозаселенные штаты;
- с самым высоким доходом населения и самые низкодоходные;
- с самым высоким уровнем образования и самые неграмотные штаты;
- штаты, где люди живут дольше и где низкая продолжительность жизни;
- самые холодные и теплые штаты;
- самые криминальные и самые спокойные штаты.

Для фиксации результата скопируйте по 3 выбранных названия штатов в пустые столбцы, например, с ячейки P5.

Определите по датасету взаимосвязи с последующими выводами:

- Какие штаты имеют высокие показатели по уровню дохода, образованности, продолжительности жизни и какие – низкие по данным параметрам.
- Какие штаты самые холодные, а какие - теплые. Зависит ли это от региона?
- В каких штатах высокий уровень убийств, а в каких - низкий. Есть ли зависимость преступности от уровня безграмотности и продолжительности жизни?

Для анализа определите коэффициенты корреляции по основным парам переменных датасета (функция КОРРЕЛ): проверьте корреляцию между климатом, образованием, продолжительностью жизни населения штатов, неграмотностью и уровнем убийств. Ответы по выводам запишите в соседних ячейках листа.

По таблице Регионы постройте столбчатую диаграмму криминальности, а по таблице Штаты (группируя по регионам) – точечную диаграмму плотности населения.

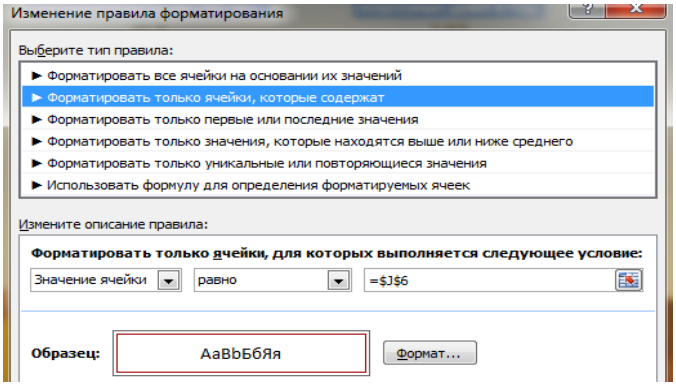
Для наглядности рассмотрения полученной информации по итогам анализа соберем на отдельном листе информационную панель (лист «дашборд») с возможностью выбора штата из списка (рисунок 5).

Для выразительности представления результатов на дашборде используйте фоновый рисунок через меню на вкладке Макет страницы в группе Параметры страницы команду Подложка. В этом примере подключен рисунок [Пустыня.jpg].

Для выбора штата из списка добавьте на лист дашборда элемент управления Поле со списком из вкладки Разработчик. Для списка названий штатов подготовьте на листе Usa_stat в колонке T сортированный диапазон и задайте ему имя statname. Свяжите элемент управления Поле с ячейкой S1 на листе Usa_stat.

Затем в ячейках S2..S13 используйте формулы =ИНДЕКС(statname;S1) | =ПОИСКПОЗ(S2;штаты[State_name];0) | =СМЕЩ(A1;S4;0) | =СМЕЩ(A1;S4;1) и т.д. – для подготовки значений, отображаемых на дашборде.

Дополнительно на лист дашборда в ячейки рядом с элементом управления Поле со списком добавьте формулы =ЕСЛИ(ЕСЛИОШИБКА(ПОИСКПОЗ('usa-stat'!S2;'usa-stat'!O3:O17;0);0)>0;"3 лучших";"") =ЕСЛИ(ЕСЛИОШИБКА(ПОИСКПОЗ('usa-stat'!S2;'usa-stat'!Q3:Q17;0);0)>0;"3 худших";"")



Для подсветки региона добавьте условное форматирование (как на этом примере).

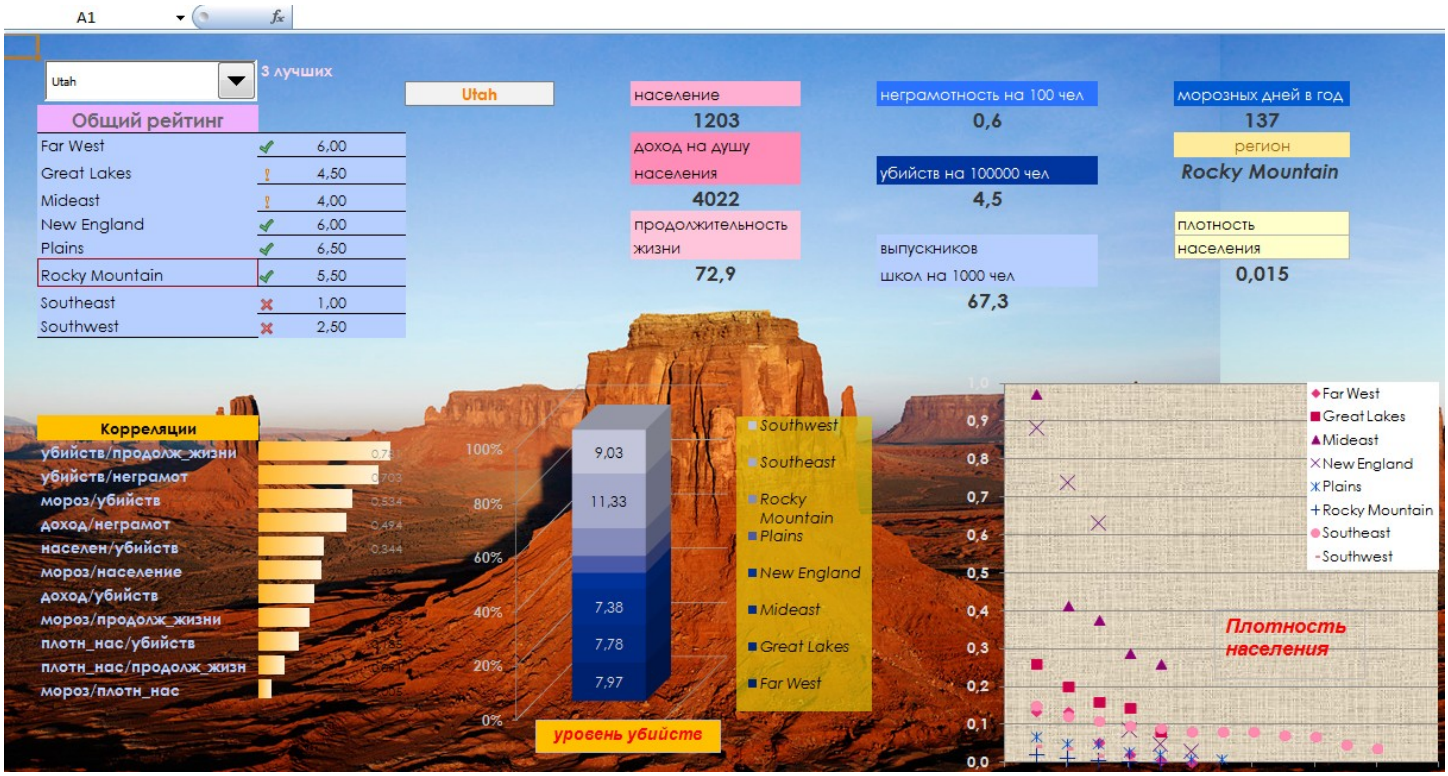


Рисунок 5 – вариант оформления дашборда

Добавьте отдельный лист, назовите его «отчет» и на нем впишите графы Фамилия Имя, Группа, Дата выполнения работы и запишите свои реквизиты.
Отправьте файл Excel на почту преподавателю.