Дипломный проект

"Анализ данных США 1970 гг. на предмет определения социального развития"

Целью исследования является:

- Определение наиболее благоприятных штатов для жизни по следующим признакам:
    - Высокий доход;
    - Образованность;
    - Благополучная криминальная обстановка;
    - Высокая продолжительность жизни;
    - Благоприятные климатические условия;
- Анализ штатов, которые испытывают проблемы с социальным развитием, очерчена их региональная принадлежность и предположительно установлены причины социальных проблем.

Таким образом, в результате анализа определена взаимосвязь факторов, влияющих на благополучие населения.

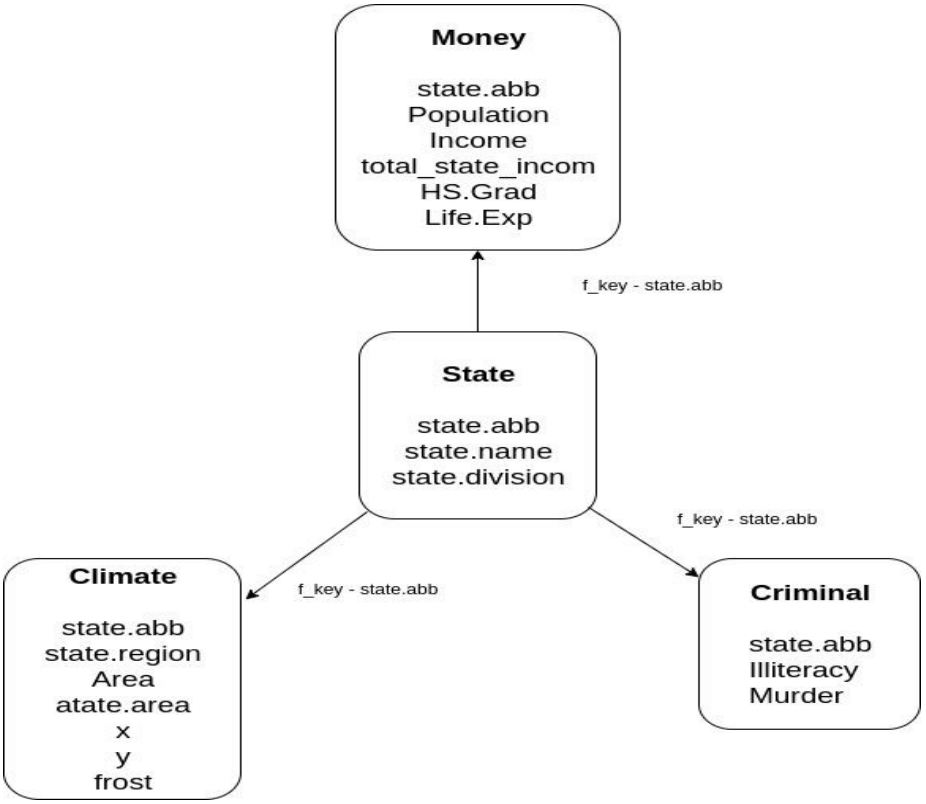
Данные для анализа были взяты отсюда. Указанный датасет содержит данные 1970-х годов по всем пятидесяти штатам США. Для каждого штата датасет включает:

- Численность населения;
- Доход на душу населения;
- Уровень неграмотности среди населения;
- Количество убийств;
- Процент выпускников средней школы;
- Среднее число морозных дней;

- Площадь, широту и долготу каждого штата;
- Региональную принадлежность и округа, к которым принадлежит тот или иной штат;
- Полное название штата и аббревиатуру.

Дополнительно я обогатила датасет, добавив метрику “Совокупный доход штата” (перемноженные метрики: численность населения на доход на душу населения). Также мной были использованы дополнительные материалы для построения дашбордов - по национальному составу США.

Для удобства анализа данные были переложены в схему “звезда”.



Я разделила данные по признакам:

- Деньги. Помимо непосредственно дохода на душу и дохода всего штата, включила и то, что непосредственно влияет на доход населения: процент выпускников средней школы, продолжительность жизни и количество населения.

- Климат. Помимо метрики по холодным дням, в набор данных включила и географические значения: широта, долгота, площадь штата, а также принадлежность штата к округу и региону.
- Преступность. Сюда включила негативные факторы: безграмотность и количество убийств. Как показало исследование, эти две метрики взаимосвязаны - чем ниже грамотность населения - тем выше количество убийств в штате (и регионе) и наоборот.
- Штат. В данных находится наименование штата: краткое и полное, а также его региональная принадлежность.

Данные всех четырех таблиц между собой связаны по ключу - state.abb - двубуквенная аббревиатура штата.

Анализ данных был проведен на Python в Jupiter Notebook и добавлен на GitHub. Также датасет был разделен и загружен в облачное хранилище BigQuery через API и БД доступна по ссылке.

После чего был проведен анализ основных показателей в Data Studio и созданы дашборды, демонстрирующие наглядность выводов, изложенных в Jupiter Notebook.

А именно:

- Влияние климата на численность населения и продолжительность жизни
- Зависимость дохода от уровня образованности
- Взаимосвязь штатов с низкой грамотностью населения и количеством тяжких преступлений

В ходе исследования по средним показателям по регионам и округам выявлено, что основные социальные проблемы в южном регионе на юго-востоке страны (East South Central) и южной атлантике (South Atlantic). Благополучный регион - север, а округа - побережье Тихого океана (Pacific), северо-запад (West North Central).

Исследовав набор данных “Money”, нашла зависимости:

1) Population к total_state_income: чем больше людей в штате, тем больше совокупный доход в штате.

2) Income к HS_Grad: чем выше процент выпускников в штате, тем больше людей продолжают учиться дальше, получают профессию и, соответственно, тем самым увеличивают доход - как свой, так и штата.

3) Life_Exp к HS_Grad: предположительная длительность жизни зависит от уровня образованности - образованный человек имеет больший доход и, соответственно, лучшее качество жизни.

А также данные по штатам:

- Самая большая численность населения - в Калифорнии;
- Самая низкая популяция - в Аляске, но при этом доход населения ($ 6315) самый высокий по стране. Аналогично дела обстоят и в Неваде: пятая по малочисленности и занимает 5 место по уровню дохода.
- В Миссисипи доход самый низкий у населения США - $ 3098.
- В Южной Каролине процент выпускников крайне мал - 37.8%. Кроме того, еще и самая низкая продолжительность жизни в стране.
- В Гавайях люди живут дольше, чем в прочих штатах. Кроме того - в Гавайях 0 морозных дней в году.
- Юта находится на 3 месте по продолжительности жизни, что говорит о высоком уровне социального развития - ведь в ней еще и самый высокий уровень образованности (67.3%).
- Аляска с Невадой, несмотря на малочисленность штата, следуют сразу за Ютой по количеству выпускников.

Исследовав набор данных “Climate”, нашла зависимость от количества холодных дней от y (широта), при этом от x (долготы) температура зависит мало. А также:

- Аляска не входит в топ самых холодных штатов (хотя у Аляски 152 дня в году холодных дней), но все же является самым северным штатом.
- Невада с высоким уровнем образования и высоким доходом на душу населения - самый холодный и малонаселенный штат.
- Гавайи - самый теплый штат, где живут дольше всего. Но прямой зависимости нет, так как в самом холодном штате - Аляске - далеко не самая низкая в США продолжительность жизни. Т.е. можно предположить, что Гавайи пользуются спросом для переезда туда пожилых обеспеченных американцев. Самая низкая продолжительность жизни - в Южной Каролине, которая является относительно теплым штатом.

После исследования набора данных “Criminal”, корреляция между неграмотностью и уровнем убийств стала очевидна. А также:

- Лидером по безграмотности является Луизиана. Дальше знакомые по антирейтингам выше - Миссисипи, Северная Каролина.
- Южная Дакота - лидер по отсутствию безграмотности, за ней следуют Невада (известная по высоким рейтингам по прочим параметрам).
- Небезопасно для жизни в Алабаме, Джорджии, Луизиане, Миссисипи, Техасе. Все 5 штатов находятся на юге страны и граничат друг с другом.
- Безопасней всего для жизни - в Северной Дакоте и прочих северных штатах.

На основании вышеизложенных фактов, можно предположить, что южные штаты серьезно отстают в социальном развитии от северных штатов. Не смотря на то, что к 1970 г. с момента окончания гражданской войны “Севера и Юга” в США (по итогу которой произошла реинтеграция проигравших южных штатов Конфедерации в состав США, а также отмена рабства) прошло чуть больше 100 лет, но отставание не удалось нивелировать - как видно из исследования. Кроме того, к 1970 г. только окончилась сегрегация в стране, со всеми вытекающими социальными проблемами - как известно, процент “белого” населения в южных штатах серьезно ниже, чем в северных.