Кластеризация

Структура экзаменационных билетов

Структура экзаменационного билета

В билете 2 задания:

- **1 задание – предварительный анализ данных** (исследование данных, визуальный анализ, фильтрация, выявление мультиколлинеарности, вывод о возможности снижения признакового пространства (корреляционный анализ, МГК (факторный анализ), кластерный анализ) – **20 баллов**
- **2 задание – прогнозирование** (построение нескольких (не больше 3) моделей регрессии или классификации и предсказание целевого признака) – **40 баллов**

Примеры задания 1

1. На основе индивидуальных данных о клиентах банка (файл “….csv”) проведите оценку основных статистических характеристик набора данных с использованием SAS Studio. Сформируйте набор данных с информацией о клиентах старше 45 лет с уровнем дохода не менее 2000 у.е. Проведите визуальный анализ полученной в результате фильтрации выборки (не менее 3 диаграмм)
2. По результатам корреляционного и компонентного/факторного анализа данных о клиентах страховой компании (файл “….csv”) сделайте вывод о возможности снижения размерности признакового пространства. Предложите смысловую интерпретацию главных компонент. Требуемый уровень информативности – не менее 80%

Примеры задания 1

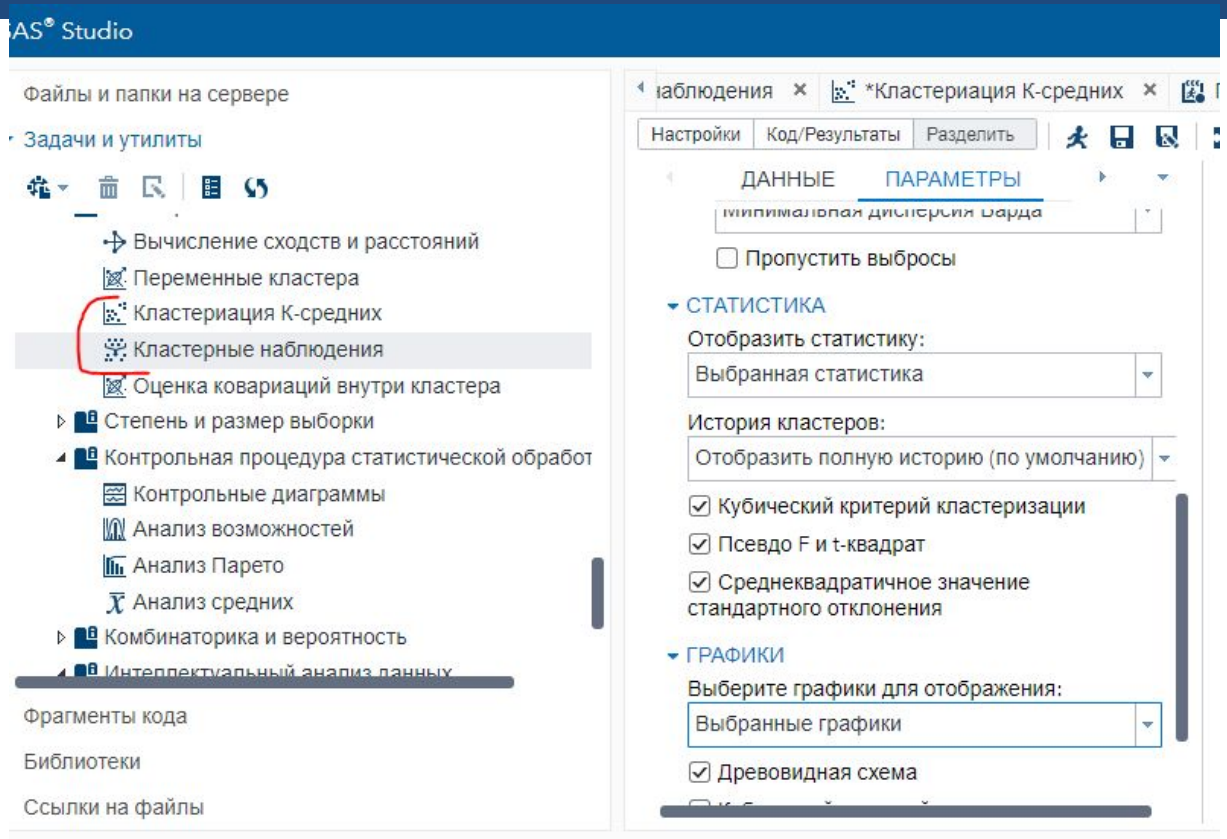
3. С использованием SAS Studio проведите кластеризацию объектов недвижимости из набора данных «...csv» на функциональные группы методом k-средних для различных вариантов настроек. Интерпретируйте полученные результаты для трех моделей с помощью отчета по кластеризации, сравните полученные результаты, сделайте выводы

Примеры задания 2

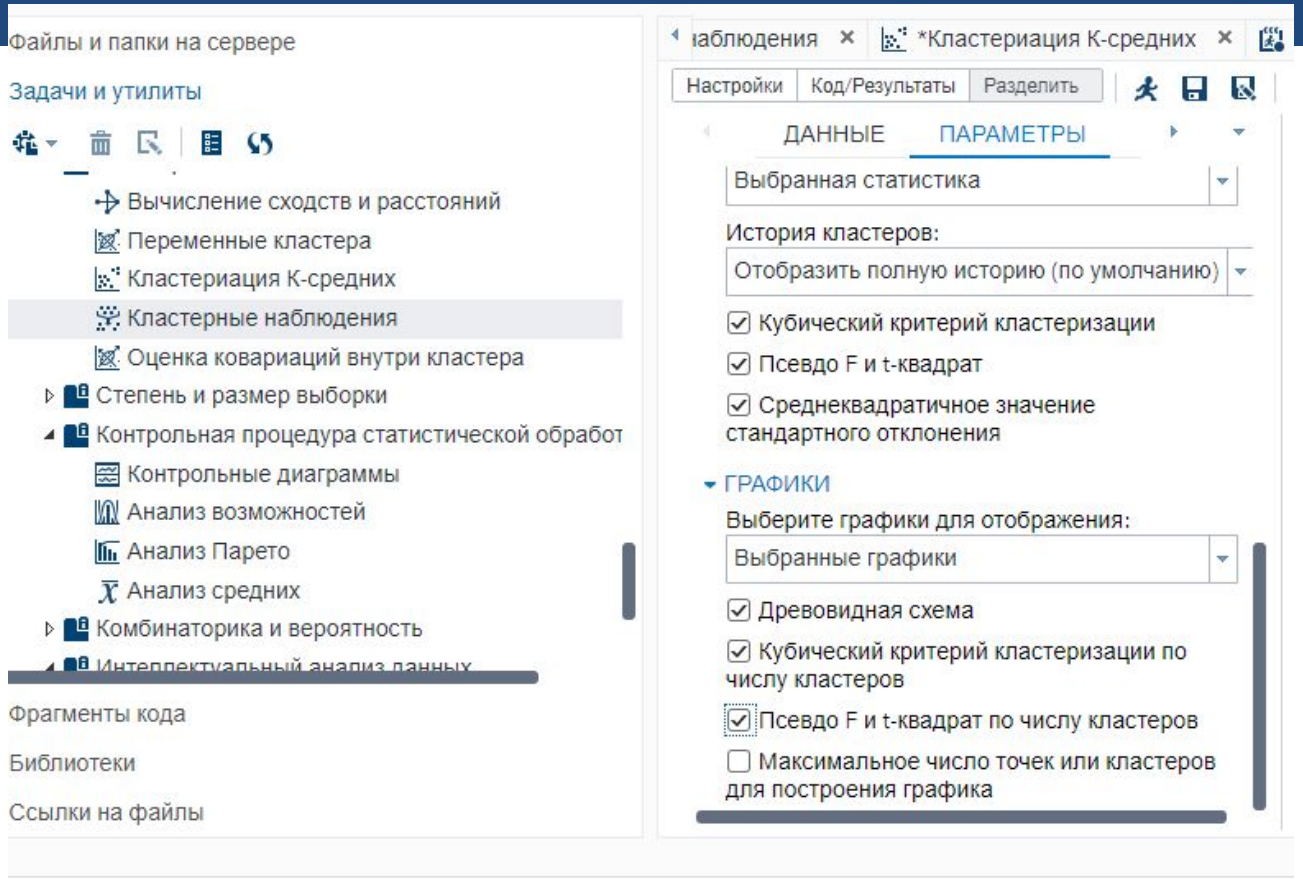
1. Постройте и исследуйте три регрессионные зависимости срока кредитования от возраста и длительности трудоустройства иностранных клиентов, арендующих жилье с использованием инструментов отбора признаков SAS/STAT. Проведите сравнительный анализ качества полученных моделей, предложите смысловую интерпретацию результатов и сделайте выводы

2. Для набора данных «...csv» построить бинарную логистическую регрессию с использованием SAS Studio (не менее 3 моделей-кандидатов, используя различные методы отбора переменных в модель и вид модели). Провести сравнение моделей кандидатов, выявить наилучшую по результатам ROC-кривой и значений показателя AUC, сделать выводы.

Кластерный анализ в SAS/STAT



Кластерный анализ в SAS/STAT

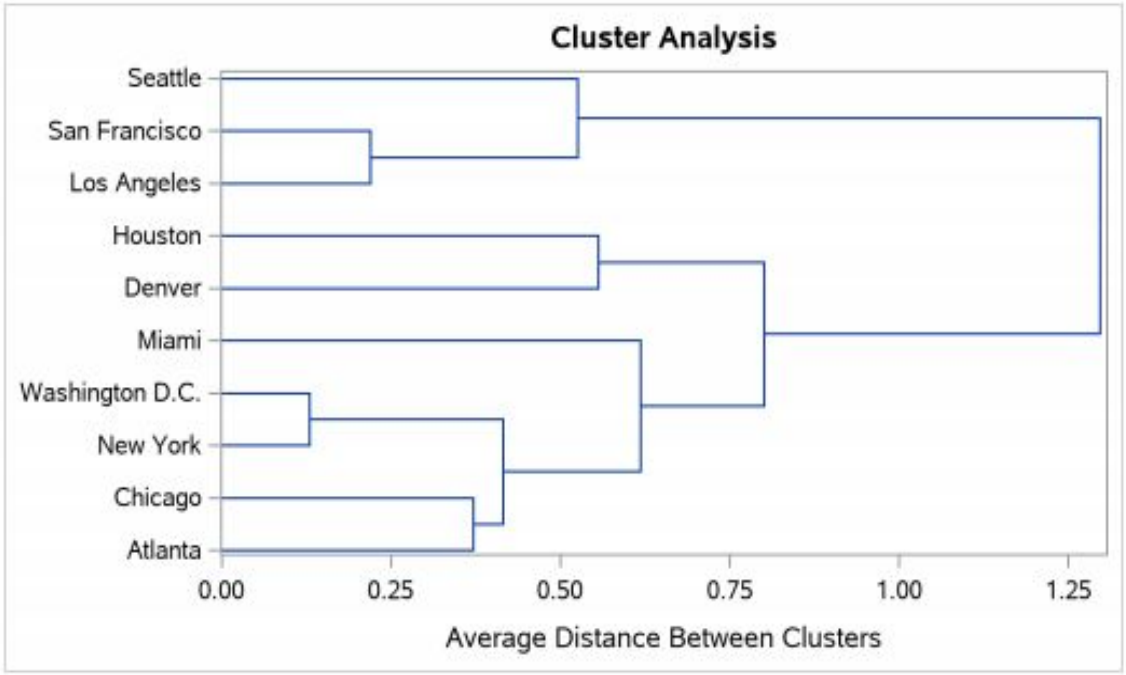


Результаты

Cluster History								
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Pseudo F Statistic	Pseudo t-Squared	Tie
9	New York	Washington D.C.	2	0.0019	.998	66.7	.	
8	Los Angeles	San Francisco	2	0.0054	.993	39.2	.	
7	Atlanta	Chicago	2	0.0153	.977	21.7	.	
6	CL7	CL9	4	0.0296	.948	14.5	3.4	
5	Denver	Houston	2	0.0344	.913	13.2	.	
4	CL8	Seattle	3	0.0391	.874	13.9	7.3	
3	CL6	Miami	5	0.0586	.816	15.5	3.8	
2	CL3	CL5	7	0.1488	.667	16.0	5.3	
1	CL2	CL4	10	0.6669	.000	.	16.0	

Дендрограмма

Output 37.1.3 Dendrogram Using METHOD=AVERAGE



Number of Clusters - количество кластеров

Clusters Joined - имена объединенных кластеров. (Наблюдения идентифицируются либо по значению идентификатора, либо по CLn, где n - номер кластера)

Freq - количество наблюдений в новом кластере

Semipartial R-Square - полупериодический квадрат R, представляет собой уменьшение доли дисперсии, приходящейся на объединение двух кластеров.

R-Square - квадратная кратная корреляция R квадрат, которая представляет собой долю дисперсии, учитываемой кластерами
Approximate Expected R-Square - примерное ожидаемое значение квадрата R. Это ожидание аппроксимируется при нулевой гипотезе о том, что данные имеют равномерное распределение вместо формирования отдельных кластеров.
В следующих трех столбцах отображаются значения статистики кубического критерия кластеризации (CCC), псевдо F (PSF) и (PST2). Эта статистика полезна для оценки количества кластеров в данных.
связи для минимального расстояния; пустое значение указывает на отсутствие связи. Связывание означает, что кластеры являются неопределенными и что изменение порядка наблюдений может изменить кластеры.

Задания

1. Выполнить задания из файла «Сем 13.10_Кластеризация.doc».

2. Выполнить кластерный анализ для набора данных из задания 2 с использованием SAS/STAT (задачи Кластеризация К-средних;

Кластерные наблюдения)