Дополнительная профессиональная программа повышения квалификации «Анализ данных: самостоятельный старт»

Программа направлена на быстрое получение базовых знаний и навыков в области анализа данных в гибком формате. В рамках обучения по программе слушатели учатся составлять аналитические отчеты, пользоваться инструментами для работы с исследовательской аналитикой, применять язык SQL для обработки данных, находить закономерности в данных с помощью их визуализации, моделировать процессы и использовать метрики и показатели в анализе данных.

№	Наименование	Характеристика
1	**Наименование и цели программы**	
1.1	Наименование	**«Анализ данных: самостоятельный старт»**
1.2	Вид программы	Дополнительная профессиональная программа повышения квалификации
1.3	Цель	Формирование у слушателей аналитического мышления и навыков обращения с данными, составления аналитических отчетов, изучение способов анализа данных, метрик и показателей, а также применение языка SQL для обработки данных.
1.4	Планируемые результаты (формируемые или совершенствующиеся компетенции в соответствии с указанным профстандартом)	Подготовка данных для проведения аналитических работ по исследованию больших данных Профессиональный стандарт 06.042 «Специалист по большим данным» (утв. приказом Министерства труда и социальной защиты РФ от 06.07.2020 № 405н). Извлечение и описание образцов данных и агрегированных значений из систем-аналогов Профессиональный стандарт 06.022 «Системный аналитик» (утв. приказом Министерства труда и социальной защиты РФ от 27.04.2023 № 367н.) Анализ, обоснование и выбор решения Профессиональный стандарт 08.037 «Бизнес-аналитик» (утв. приказом Министерства труда и социальной защиты РФ от 22.11.2023 № 821н.)
2	**Формат реализации**	
2.1	Язык обучения	Русский
2.2	Форма обучения	Заочная с применением ДОТ

2.3	Объем программы	30 академических часов
2.4	Режим занятий	В течение 4 недель в удобное для слушателей время
2.5	Выдаваемый документ	Удостоверение о повышении квалификации Академии ПСБ
2.6	Целевая аудитория	Аналитики данных; руководители проектов и менеджеры в области информационных технологий и аналитики; студенты и начинающие специалисты по анализу данных; лица, интересующиеся анализом данных и желающие приобрести базовые знания и навыки в области анализа данных
2.7	Численность обучаемых по программе	Численность слушателей определяется договором об образовании, заключаемым при приеме на обучение
3	**Организационно-педагогические условия**	
3.1	Учебный план	см. Приложение № 1
3.2	Календарный учебный график	см. Приложение № 2
3.3	Рабочая программа (учебные предметы, курсы, дисциплины (модули)	см. Приложение № 3
3.4	Профессорско-преподавательский состав	Реализация программы обеспечивается квалифицированными научно-педагогическими кадрами, имеющими базовое образование, соответствующее профилю преподаваемой дисциплины и систематически занимающимися научной и научно-методической деятельностью
3.5	Форма итоговой аттестации	Экзамен в форме выполнения практического задания
3.6	Оценочные материалы	Пример задания см. Приложение № 4. Результаты выполнения задания оцениваются на основании 4 критериев, с присвоением баллов по каждому критерию оценки. Максимальный балл, который может набрать слушатель за выполнение итогового задания – 100 баллов. Условием успешного завершения обучения и получения удостоверения о повышении квалификации является выполнение итогового задания с результатом не менее 40 баллов
3.7	Методические материалы	Учебная литература (учебники, пособия, книги) – предоставляется доступ к электронной библиотечной системе «Знаниум» (www.znanium.ru).

4	**Контактная информация**	
4.1	Руководитель программы	Мубаракшина Алсу Ирековна, директор Центра развития обучения по управлению данными АНО ДПО «Академия ПСБ» a.mubarakshina@psb-academy.ru

Приложение № 1

УЧЕБНЫЙ ПЛАН
дополнительной профессиональной программы повышения квалификации
«Аналитик данных: самостоятельный старт»

Количество часов – 30 академических часов.

Форма обучения – заочная с применением дистанционных образовательных технологий.

№ п/п	Наименование раздела/ модуля, темы	Общая трудоемкость, час.	Аудиторная работа, час.					С применением дистанционных образовательных технологий, электронного обучения, час.					Самостоятельная работа, час.	Текущий контроль успеваемости	Промежуточная аттестация (форма/час.)	Итоговая аттестация (вид/час.)	Код компетенции
			Всего	В форме практической подготовки	В том числе			Всего	В форме практической подготовки	В том числе							
					Лекции/в интерактивной форме	Практические (семинарские) занятия/в интерактивной форме	Контактная самостоятельная работа			Лекции/в интерактивной форме	Практические (семинарские) занятия/в интерактивной форме	Контактная самостоятельная работа					
1	**Введение**	**2**						1,5	1	0,5		1		0,5			A/03.6
2	**Аналитическое мышление**	**1**						1	0,5	0,5		0,5					D/02.6
3	**Визуализация**	**2,5**						2,5	0,5	2		0,5					D/02.6
4	**Аналитический конвейер**	**1,5**						1,5	0,5	1		0,5					D/02.6
5	**Консолидация данных**	**2,5**						2,5	0,5	2		0,5					A/03.6
6	**SQL**	**14**						14	0,5	13,5		0,5					A/04.4
7	**Моделирование**	**1**						1	0,5	0,5		0,5					A/04.4
8	**Аналитический продукт**	**1**						1	0,5	0,5		0,5					A/03.6

9	**Машинное обучение**	**2**						2	0,5	1,5		0,5					A/03.6
10	**Инженерия данных**	**1,5**						1,5	0,5	1		0,5					A/03.6
	Итоговая аттестация	**1**														**Э[1]/1**	A/03.6, D/02.6, A/04.4
	ИТОГО:	**30**						**28,5**	**5,5**	**23**		**5,5**		**0,5**		**1**	

[1] Вид итоговой аттестации: Э – экзамен.

Приложение № 2

КАЛЕНДАРНЫЙ УЧЕБНЫЙ ГРАФИК

Период обучения:		
30 ак. часов	4 недели	1 месяц

РАБОЧИЕ ПРОГРАММЫ РАЗДЕЛОВ / МОДУЛЕЙ / ДИСЦИПЛИН

1. ВВЕДЕНИЕ
1.1. Входное тестирование. 1.2. Введение в анализ данных. • Определение аналитика данных. • Задачи аналитика. • Осознанность и осмысленность.
2. АНАЛИТИЧЕСКОЕ МЫШЛЕНИЕ
2.1. Аналитическое мышление. • Постановка задачи. • Формальные отчеты. • Исследовательская аналитика. • Задачи и инструменты.
3. ВИЗУАЛИЗАЦИЯ
3.1. Визуализация в анализе данных. • Польза и удобство визуализации для аналитика. 3.2. Выбор способа визуализации. • Выбор подходящего способа визуализации для разных типов задач. • Описание показателей (мер) и факторов (категорий) в таблице.
3.3. Задачи визуализации. • Зависимость вида визуализации от её задачи.
3.4. Подписи и заголовки. • Важность написания подходящей подписи, заголовка, подбора иллюстрации.
4. АНАЛИТИЧЕСКИЙ КОНВЕЙЕР
4.1. Аналитический конвейер. • Область проблем пользователя. • Пользовательские требования.
5. КОНСОЛИДАЦИЯ ДАННЫХ
5.1. Консолидация данных. • Определение консолидации данных. • Задачи консолидации данных. • Источники данных.
5.2. Консолидация в корпоративной среде. • Корпоративное хранилище данных. • Преобразование данных.
5.3. Идеи для консолидации данных. • Единое хранилище данных. • События и контекст событий. • Конвейер для загрузки данных.
5.4. Пример консолидации данных. • Пример решения задач по консолидации данных.
6. SQL
6.1. SQL. • Понимание реляционных баз данных.

6.2. Клиент-серверная архитектура. • Описание файл-серверной и клиент-серверной архитектуры.
6.3. Столбцы и строки. • Столбцы как структура данных. • 4 основные инструкции для пользователя при работе со строками.
6.4. Понимание таблиц. • Таблицы связей.
6.5. Инструкция SELECT. • Описание работы секций SELECT и FROM.
6.6. Фильтрация и выборка. • Фильтрация по столбцам. • Фильтрация по строкам.
6.7. Использование функций. • Применение функций в запросе.
6.8. Неизвестные значения. • Недостаток реляционных баз данных. • Способы устранения недостатков реляционных баз данных.
6.9. Сортировка. • Способы управления порядком передачи результата.
6.10. Отсечка. • Понятие «отсечка» и примеры применения.
6.11. Устранение дубликатов. • Как работать с дубликатами.
6.12. Порядок операций в запросе. • Важность порядка операций в запросе.
6.13. Трансформирующие операции. • Знакомство с трансформирующими операциями. • Понимание чем является каждая строка.
6.14. Агрегация. • Устройство агрегатных функций. • Различие агрегатных и скалярных функций.
6.15. Группировка. • Важность использования группировки. • Практическое применение группировки.
6.16. Фильтрация групп. • Операция HAVING. • Примеры решения задач.
7. МОДЕЛИРОВАНИЕ
7.1. Моделирование в анализе данных. • Задачи моделирования. • Метрики, показатели, KPI. • Требования к модели данных.
8. АНАЛИТИЧЕСКИЙ ПРОДУКТ
8.1. Аналитический продукт. • Задачи отчета. • Готовый аналитический продукт. • Форматы готового аналитического продукта.
9. МАШИННОЕ ОБУЧЕНИЕ

9.1. Датасайнс. Класс задач для применения датасайнс.
9.2. Машинное обучение. • Определение машинного обучения. • Конвейер машинного обучения.
10. ИНЖЕНЕРИЯ ДАННЫХ
10.1. Инженерия данных. • Разработка ETL-процессов. • Задачи дата-инженера.
10.2. Большие данные. • Отслеживание жизненного цикла данных.

Пример задания итоговой аттестации

Итоговое задание состоит из двух задач, которые соответствуют этапам «аналитического конвейера».

Задача №1. Привести к удобному формату.

Исходная таблица довольно большая. Предложите способ (напишите SQL-запрос), который позволит радикально уменьшить число строк.

Ваш запрос должен обязательно сохранить имеющиеся в данных закономерности. То есть, если есть какие-то заметные особенности в данных и эти особенности могут быть интересны в рамках нашего аналитического сценария, то эти закономерности должны сохраниться и в сокращённом наборе данных.

Смысл задания заключается в том, чтобы провести анализ на сокращённом наборе данных и прийти к тем же выводам, как при анализе исходной большой таблицы.

Задача № 2. Провести аналитическую обработку.

Найдите в данных какую-либо закономерность, интересную в рамках нашего аналитического сценария и покажите эту закономерность наглядно при помощи SQL-запроса. Запрос должен выдавать результат, который будет сразу понятен без какой-либо дополнительной визуализации (без условного форматирования, без графиков и прочего).

Сама таблица, которую выдаст ваш SELECT, должна иллюстрировать найденную закономерность.

Не обязательно искать сложную и неочевидную закономерность. Суть этого задания состоит в осмысленном применении фильтрации, группировки, агрегации, сортировки – тех операций, которые изучались для аналитической задачи.

Описание задания:

На учебном сервере SQL.DataExplorer.ru найдите базу данных "(хакатон Azure ML в Микрософте) 21-22 мая 2016". В этой базе данных исследуйте таблицу "Мать и дитя"."MS_Data05".

Она содержит обезличенные данные по процедурам, которые проводились для пациентов сети клиник "Мать и дитя" в 2014 – 2015 годах.

Каждая строка этой таблицы – это обслуживание одного клиента по одной процедуре в течение одного дня. Если клиенту потребовалось несколько одинаковых процедур в один день, то столбец SERV_COUNT содержит значение больше единицы.

Уникальный номер пациента	Дата рождения и пол пациента		Дата проведения процедуры	Код процедуры и код типа процедур		Количество процедур и выручка		Номер отделения и регион	
PATIENTS_ID	PATIENTS_BIRTHDAY	PATIENTS_POL	DATE_EXEC	SERV_ID	SERV_GROUP_ID	SERV_COUNT	SERV_SUM	FILIAL_ID	REGION_ID
111752	2006-04-25	0	2014-09-20	20229	2028	1	2560,00	1	1
321330	1987-02-02	1	2014-07-29	31716	2400	1	720,00	116011	9
320916	1983-02-17	1	2014-09-27	31814	2642	1	17920,00	116011	9
178049	2005-11-10	1	2014-07-01	26122	2391	2	3040,00	81312	3
390770	1980-06-26	1	2015-07-16	31357	2026	1	680,00	126836	5

Рисунок 1. Пример из базы данных "(хакатон Azure ML в Микрософте) 21-22 мая 2016"

Вопрос по заданию:

Оценить половозрастной состав пациентов в разных регионах.

Решение задания:

Задача №1.

Идеи для сокращения количества строк:

1) Поскольку в задаче не требуется детализация по медицинским процедурам, можно сгруппировать таблицу, исключив лишнюю детализацию:

```
ELECT                   DATE_EXEC,
                                PATIENTS_POL,
                                PATIENTS_BIRTHDAY,
                                REGION_ID,
                                FILIAL_ID,
                                Sum (SERV_COUNT) AS SERV_COUNT,
                                Sum (SERV_SUM) AS SERV_SUM
FROM                    "Мать и дитя"."MS_Data05"
GROUP BY        DATE_EXEC,
                                PATIENTS_POL,
                                PATIENTS_BIRTHDAY,
                                REGION_ID,
                                FILIAL_ID
```

2) Можно укрупнить периоды до месяцев:

```
SELECT                  Year (DATE_EXEC) AS "Year",
                                Month (DATE_EXEC) AS "Month",
                                SERV_GROUP_ID,
                                SERV_ID,
                                REGION_ID,
                                FILIAL_ID,
                                PATIENTS_ID,
                                PATIENTS_POL,
                                PATIENTS_BIRTHDAY,
                                SERV_ID,
                                Sum (SERV_COUNT) AS SERV_COUNT,
                                Sum (SERV_SUM) AS SERV_SUM
FROM                    "Мать и дитя"."MS_Data05"
GROUP BY        Year (DATE_EXEC),
                                Month (DATE_EXEC),
                                SERV_GROUP_ID,
                                SERV_ID,
                                REGION_ID,
                                FILIAL_ID,
                                PATIENTS_ID,
                                PATIENTS_POL,
```

```
                                PATIENTS_BIRTHDAY,
                                SERV_ID
```

3) Года рождения пациента достаточно для того, чтобы оценить возраст. Можно "укрупнить" пациентов до групп "пол – год рождения":

```
ELECT           DATE_EXEC,
                        SERV_GROUP_ID,
                        SERV_ID,
                        REGION_ID,
                        FILIAL_ID,
                        PATIENTS_POL,
                        Year (PATIENTS_BIRTHDAY) AS
Birthday_Year,
                        Sum (SERV_COUNT) AS SERV_COUNT,
                        Sum (SERV_SUM) AS SERV_SUM
FROM            "Мать и дитя"."MS_Data05"
GROUP BY        DATE_EXEC,
                        SERV_GROUP_ID,
                        SERV_ID,
                        REGION_ID,
                        FILIAL_ID,
                        PATIENTS_POL,
                        Year (PATIENTS_BIRTHDAY)
```

4) Совместить эти три подхода – сгруппировать, убрать лишние столбцы и укрупнить периоды:

```
SELECT          Year (DATE_EXEC),
                        Month (DATE_EXEC),
                        SERV_GROUP_ID,
                        REGION_ID,
                        PATIENTS_POL,
                        Year (PATIENTS_BIRTHDAY) AS
Birthday_Year,
                        Sum (SERV_COUNT) AS SERV_COUNT,
                        Sum (SERV_SUM) AS SERV_SUM
FROM            "Мать и дитя"."MS_Data05"
GROUP BY        Year (DATE_EXEC),
                        Month (DATE_EXEC),
                        SERV_GROUP_ID,
                        REGION_ID,
                        PATIENTS_POL,
                        Year (PATIENTS_BIRTHDAY)
```

Задача №2.

Пример оценки половозрастного состава пациентов по регионам.

Среди женщин самые молодые пациенты (в среднем) – в регионе №3 и с заметным отрывом.

```
SELECT          REGION_ID,
                        Avg (Year (PATIENTS_BIRTHDAY)) AS "Год рождения",
                        Sum (SERV_SUM) AS Выручка
FROM            "Мать и дитя"."MS_Data05"
WHERE           PATIENTS_POL = 1
GROUP BY        REGION_ID
ORDER BY        Avg (Year (PATIENTS_BIRTHDAY)) DESC
```

Среди клиентов женщин в три раза больше, чем мужчин.

```
SELECT          PATIENTS_POL,
                        Count (DISTINCT PATIENTS_ID) AS Пациентов,
                        Sum (SERV_SUM) AS Выручка
FROM            "Мать и дитя"."MS_Data05"
WHERE           PATIENTS_POL IN (0, 1)
GROUP BY        PATIENTS_POL
```