https://codeload.github.com/LiutongZhou/NYC-Taxi-Trip-Carpooling-Assistant/zip/2.1

مجموعه داده های مورد استفاده در این پروژه از وب سایت دولتی NYC دریافت شده است. در این سایت مجموعه داده های سفرها به تفکیک سال موجود است که حجم فایل ها حداقل 1 GB می باشد و بالغ بر 9 میلیون رکورد را شامل میشود. با توجه به حجم زیاد دیتا، تنها به دیتای ماه ژانویه سال 2016 میلادی بسنده شده است. 157300 رکورد دیتای ارزشمند برای انجام پروژه مدکور استخراج شد که در نهایت هدف از پردازش این داده ها، ادغام سفرها جهت بهینه سازی تردد در نیویورک سیتی به مقصد JFK Airport می باشد.

داده ها به ازای هر سفر شامل مقادیر به شرح ذیل می باشد:

پارامترهای طول و عرض جغرافیایی مسافران، محدوده JFK، و رانندگان

تعداد مسافر

زبان مسافر

وضعیت سیگاری بودن

سمت شغلی مسافر

وضعیت تاهل مسافر

سرعت ماشین

تاریخ سفر

ساعت شروع و پایان سفر

میزان تاخیر در سفر

طول مدت سفر.

که در این پروژه، مباحث مرتبط با امنیت سفر ، کیفیت و دغدغه های مسافران و رانندگان در نظر گرفته نشده است لذا پارامترهای مرتبط با این مباحت در پردازش دیتا موثر واقع نشده اند.

در جریان پیش پردازش داده ها سه جدول در پایگاه داده ایجاد می شود.

این جداول شامل سفرها، سفرهای ادغام شده و جدول تقاطع ها است.

در نهایت نمایشی از سفرهای ادغام شده خواهیم داشت.

معیارهای ادغام سفرها :

مسافران خواهان سفارش سفر از نقاط مبدا (حداکثر شعاع 100 متری یکدیگر واقع شده باشند) به نقاط مقصد حداکثر دارای سه تقاطع مشترک در ناحیه JFK هستند. نقاط مشترک در طول مسیر نهایتا به ازای یک مسافر در نظر گرفته می شوند لذا پارامتر تاخیر در سفر نیز گنجانده شده است.

تعداد مسافران حداکثر سه نفر می باشد.